UNIVERSIDADE DE BRASÍLIA (UNB)
FACULDADE DE CIÊNCIA DA INFORMAÇÃO (FCI)
CURSO DE GRADUAÇÃO EM BIBLIOTECONOMIA

Análise do perfil do profissional da informação para a atuação como cientista de dados em ambientes de big data: uma perspectiva a partir das disciplinas do curso de biblioteconomia da UnB

Mesaque Vidal Rolim

Brasília
2018

Mesaque Vidal Rolim

Análise do perfil do profissional da informação para a atuação como cientista de dados em ambientes de big data: uma perspectiva a partir das disciplinas do curso de biblioteconomia da UnB

Monografia apresentada à banca examinadora como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia pela Faculdade de Ciência da Informação da Universidade de Brasília.

Orientador: Dr. Márcio de Carvalho Victorino

Brasília
2018

R748a Rolim, Mesaque Vidal, 1991-

Análise do perfil do profissional da informação para a atuação como cientista de dados em ambientes de big data : uma perspectiva a partir das disciplinas do curso de biblioteconomia da UnB / Mesaque Vidal Rolim; Orientador: Dr. Márcio de Carvalho Victorino. – Brasília, 2018.

68 f. : il.

Monografia (Bacharelado em Biblioteconomia) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2018.

Bibliografia: p. 62-68.

1. Big data. 2. Profissional da informação. 3 Ciência da informação. 4. Ciência de dados. I. Victorino, Márcio de Carvalho, orient. II. Título.

CDU 001.102-051:004

Titulo: Análise do perfil do profissional da informação para a atuação como cientista de dados em ambientes de big data: uma perspectiva a partir das disciplinas do curso de biblioteconomia da UnB.

Aluno: Mesaque Vidal Rolim.

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília, como parte dos requisitos para obtenção do grau de Bacharel em Biblioteconomia.

Brasília, 03 de julho de 2018.

Márcio de Carvalho Victorino - Orientador
Professor da Faculdade de Ciência da Informação (UnB)
Doutor em Ciência da Informação

João de Melo Maricato – Membro
Professor da Faculdade de Ciência da Informação (UnB)
Doutor em Ciência da Informação

Dalton Lopes Martins – Membro
Professor da Faculdade de Ciência da informação (UnB)
Doutor em Ciências da Informação

Dedico este trabalho a familiares e amigos que, ao longo desses 27 anos, contribuíram de alguma maneira com a minha formação pessoal e social.

AGRADECIMENTOS

Agradeço, primeiramente, à minha mãe, Antônia, e ao senhor Ribamar por terem incentivado e assegurado minha permanência na vida acadêmica. Se hoje consigo chegar neste momento, muito se deve a vocês.

Agradeço profundamente ao meu orientador, Prof. Dr. Márcio de Carvalho Victorino, por sua disposição em auxiliar na construção deste trabalho. Suas intervenções foram de grande valia para o desenvolvimento e êxito desta pesquisa.

Agradeço a todos os meus amigos e colegas de curso que ao longo desses quatro anos contribuíram direta e indiretamente com a minha formação profissional. Agradeço, de maneira especial, às amigas Cristina Fonseca e Paola Polo pelos trabalhos produzidos em conjunto, dando sempre mostras de grande comprometimento e responsabilidade.

Agradeço às instituições onde desenvolvi atividades de estágio, Biblioteca da Escola Superior do Ministério Público da União e Coordenação de Edições Técnicas do Senado Federal, por terem acrescido competências pessoais e profissionais que hoje mostram-se significativas para a elaboração deste trabalho. Um agradecimento especial ao meu antigo supervisor de estágio, Artur Virginio dos Santos, pelos conselhos e ensinamentos compartilhados que levarei para a vida.

Agradeço à Patrícia Martins e Bárbara Queiroz, pessoas queridas que em momentos de questionamento e insegurança sempre tiveram palavras de encorajamento a oferecer.

Agradeço à Adriane Rangel pela cortesia de ter revisado o *Abstract* desta pesquisa e pela confiança e positividade transmitida através de nossas conversas.

Aos Profs. Drs. Carla Peixoto Borges, Dalton Lopes Martins, Edison Ishikawa, José Angelo Belloni, Márcio Bezerra da Silva, Maristela Terto de Holanda e, novamente, Márcio de Carvalho Victorino, que contribuíram com esta monografia disponibilizando os planos de ensino de disciplinas que ministram, meu agradecimento.

Agradeço a todos os funcionários da Faculdade de Ciência da Informação (FCI) por seu empenho em propiciar, apesar das adversidades, um bom ambiente de ensino-aprendizagem. Agradeço, especialmente, a todos os docentes com os quais tive o prazer de interagir ao longo dessa jornada, muito deste trabalho é creditado a vocês.

O gênio, esse poder que deslumbra os olhos humanos, não é outra coisa senão a perseverança bem disfarçada.

Johann Goethe

RESUMO

O trabalho busca contextualizar e conceituar a área de ciência da informação, ciência de dados, bem como o ambiente *big data*. O estudo tem como objetivo apresentar o perfil do profissional da informação, tal e qual, as contribuições que este pode trazer ao ambiente de grande volume de dados e à ciência de dados. Como metodologia é utilizada a pesquisa bibliográfica em conjunto com uma análise de currículo do curso de biblioteconomia da Universidade de Brasília (UnB), com intuito de identificar as competências esperadas do profissional da informação e do cientista de dados, assim como se há preparação do profissional da informação, em princípios básicos, para poder atuar no mercado de ciência de dados. Os resultados são apresentados por meio de um quadro comparativo entre as habilidades esperadas de um cientista de dados e aquelas previstas no profissional da informação. Além disso, são apontadas as disciplinas do curso da UnB identificadas como relevantes para atuação do futuro profissional da informação como cientista de dados, bem como a associação dessas disciplinas com o conjunto de habilidades almejadas no cientista de dados. Em conclusão, verifica-se haver espaço para o profissional da informação na área de ciência de dados, mesmo sendo um campo visto, em predominância, por meio da ciência da computação e estatística. De mais a mais, constata-se existir, no curso de biblioteconomia da UnB, um conjunto de disciplinas que permitem o desenvolvimento de competências basilares para a compreensão da área de ciência de dados.

Palavras-chave: *Big data*. Ciência da informação. Profissional da informação. Ciência de dados. Cientista de dados.

ABSTRACT

The purpose of this research is to contextualize and to conceptualize the information science area, data science and big data environment. This study aims to present the profile of the information professional, as well as their contributions to the environment of large volume data and data science. As methodology, bibliographic research was used in conjunction with a curriculum analysis of the librarianship course of University of Brasília (UnB) in order to identify expected competences of the information professional and the data scientist, as well as if there is preparation of those professionals, in basic principles, to be able to act in the data science market. The results are presented through a comparative table between the expected abilities of a data scientist and those predicted in the information professional. In addition, it is presented the disciplines of the UnB course identified as relevant for the performance of the future information professional as a data scientist, as well as the association of these disciplines with the set of skills required in a data scientist. In conclusion, there is space for the information professional in data science area, despite the fact that this is a field predominantly explored by computer science and statistics. Furthermore, it is verified that there is, in the UnB course of librarianship, a set of disciplines that allows the basic skills development for comprehension of data science area.

Keywords: Big data. Information science. Information professional. Data science. Data scientist.

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

BI *Business Intelligence*

CI Ciência da Informação

MIP Moderno Profissional da Informação

TI Tecnologia da Informação

UnB Universidade de Brasília

SUMÁRIO

1 INTRODUÇÃO

A grande evolução tecnológica ocorrida nos últimos anos, e que está em constante progresso, vem impactando diretamente diversos segmentos da sociedade moderna, e com a ciência da informação (CI) não seria diferente. Um dos efeitos mais perceptíveis dessa constatação é como o cotidiano dos indivíduos passa a estar repleto de dados e informações de variadas origens. Atividades do dia a dia das pessoas, que antes não poderiam ser monitoradas devido a limitação tecnológica, passaram a ser fontes importantíssimas para a obtenção de dados e, consequentemente, de informação. Registros do cotidiano como o desempenho da educação, questões de saúde, bens e serviços, fatores relacionados ao Estado, estatísticas sobre a economia, dados sobre o consumismo etc., "passam a nos ajudar a tomar decisões e gerar conhecimento" (RIBEIRO, 2014, p. 98).

A realidade descrita acima pode ser entendida em grande parte pela dinâmica social vivida nos dias atuais. O uso de dispositivos móveis (*smartphones*, *tablets*, *laptops*), sensores residenciais e industriais, o compartilhamento de fotos, vídeos e *e-mails*, a utilização constante de redes sociais, assim como de aplicativos, o comércio eletrônico (*e-commerce*), a geolocalização por meio de GPS, tudo isso vem produzindo quantidades torrenciais de dados.

Com esse cenário de grande profusão de dados acaba surgindo a indagação de como isso impacta áreas como a CI, assim como o que se espera do profissional que lida diuturnamente com os processos de gerar, selecionar, representar, armazenar, recuperar, distribuir e usar a informação, tendo em vista esse ambiente atual de dados sendo produzidos a todo instante.

A bem da verdade, o profissional da informação já se viu frente a frente com um desafio bastante similar a este enfrentado atualmente. Com o fim da Segunda Guerra Mundial, alguns estudiosos identificaram uma crescente proliferação de informação que estava sendo produzida, principalmente em relação às áreas de ciência e tecnologia. Ao mesmo tempo verificou-se o problema que essa explosão informacional trazia, posto que, se por um lado havia uma produção cada vez maior de informação, por outro, também havia maior dificuldade para ter acesso a essa informação gerada. Apoiado nesse contexto, um importante cientista norte-americano, Vannevar Bush, escreve sobre a utilização das insipientes tecnologias da época com o intuito de preservar e disseminar o conhecimento que estava sendo produzido, chegando a propor a construção de uma máquina (MEMEX) para esse fim (SARACEVIC, 1996, p. 42).

Hoje a situação pode ser interpretada com muita semelhança aquela verificada nos meados da década de 1940. Percebe-se aqui o surgimento de uma nova conjuntura, a geração de dados em quantidades extensas, nos mais diversos formatos e com rapidez impressionante, o *big data*. Da mesma maneira que no passado os pesquisadores tiveram que ocupar-se do problema da explosão informacional, propondo soluções para organizar e facilitar a recuperação das informações pertinentes, principalmente para o desenvolvimento científico, hoje faz-se necessário que os profissionais, das mais diversas áreas, busquem novas formas para o manejo e tratamento dessa avalanche de dados produzidos em ambientes digitais, com o intuito de extrair informação significativa para diversos fins, emergindo, desta maneira, a ciência de dados.

A partir do cenário exposto, este trabalho apresenta, inicialmente, referencial teórico contextualizando as áreas de ciência da informação e ciência de dados, tal qual, o ambiente *big data*. Além disso, delimita as características e competências atribuídas ao cientista de dados, assim como ao profissional da informação. O estudo utiliza como metodologia de trabalho a pesquisa bibliográfica, com intuito de estabelecer as habilidades e competências que contornam o cientista de dados e o profissional da informação, tal como, uma análise de currículo do curso de biblioteconomia da Universidade de Brasília, objetivando identificar disciplinas que corroboram com a atuação do profissional da informação na seara do cientista de dados. Como resultado, apresenta uma intersecção entre as habilidades dos dois profissionais estudados, assim como breve descrição das disciplinas identificadas no currículo do curso da UnB. Por fim, o trabalho discorre sobre conclusões encontradas a partir da pesquisa.

2 PROBLEMA DE PESQUISA

Como o surgimento desse novo cenário de grande profusão de dados, com grande variedade e velocidade, e, consequentemente, da ciência de dados impacta a formação do bibliotecário e a atuação do profissional da informação?

2.1 Justificativa

Desde o seu surgimento, a CI mostra-se uma área dotada de características de ciência interdisciplinar. Os campos que contribuíram com o desenvolvimento dessa disciplina são os mais diversos, perpassando a psicologia, linguística, comunicação, estatística, matemática, entre outros. Todavia, a área que contribuiu de maneira significativa com a evolução da CI foi a ciência da computação. O fenômeno observado, em meados da década de 1940, de enorme profusão informacional, essencial para o avanço da CI, se deu, em grande parte, ao desenvolvimento tecnológico observado nessa mesma época, alavancado, principalmente, pelas primeiras experiências que originaram a computação moderna (SARACEVIC, 1996; LE COADIC, 1996).

A ciência da computação foi uma das principais disciplinas colaboradoras para o progresso da ciência da informação, como já mencionado anteriormente. A ciência da computação, com o passar do tempo, foi aperfeiçoando-se gradativamente, contribuindo, assim, para a criação de novas tecnologias que, por sua vez, colaboraram com o surgimento de novos conceitos e paradigmas computacionais. Nessa perspectiva, o *big data* é um acontecimento que está diretamente ligado aos avanços tecnológicos, sobretudo ao surgimento e desenvolvimento da internet, assim como de sua maior universalização. Esse fenômeno se caracteriza pela grande quantidade de dados produzidos em ambiente virtual, tal como pela velocidade necessária ao processamento e pela variedade de formatos desses dados que são produzidos, podendo ser dados estruturados, semiestruturados e não estruturados.

Concomitantemente ao advento do *big data*, surge também uma nova ciência, a chamada ciência de dados. Essa área passa a ser caracterizada pela busca de novas técnicas e metodologias que colaborem com o armazenamento, tratamento e uso desses dados produzidos em larga escala. Da mesma maneira que a ciência da informação, a ciência de dados é qualificada como uma área interdisciplinar, sendo fundada, mormente, em princípios vindos da ciência da computação e da estatística.

Fundamentado no cenário apresentado, este estudo se justifica, sobretudo, pela importância da ciência da informação se manter atualizada junto aos novos paradigmas tecnológicos que emergem, em especial aqueles com incrível potencial de influenciar, de maneira direta, a área. Um conceito amplamente transmitido na CI é o de que a partir do dado se produz informação, posto que, quando um dado passa a ser inserido dentro de um contexto, ele obtém significado, gerando, consequentemente, informação. Baseando-se nesse princípio, e entendendo que a razão de ser da disciplina CI é o estudo de como a informação é produzida, armazenada, disseminada e utilizada, torna-se indispensável aos profissionais desse campo compreender o fenômeno de grande proliferação de dados em ambiente virtual (*big data*), assim como de que forma essa nova realidade impacta sua formação e atuação profissional, e que papel esses profissionais da informação podem desempenhar dentro dessa nova disciplina que emerge, a ciência de dados.

3 OBJETIVOS

3.1 Objetivo Geral

- Apresentar o perfil do profissional da informação, juntamente com as contribuições que este pode trazer para o ambiente *big data* e a ciência de dados, a partir de estudo de literatura da área e do currículo do curso de biblioteconomia da UnB.

3.2 Objetivos Específicos

- Conceituar e contextualizar a área de ciência da informação, assim como as competências do profissional da informação;
- Conceituar e contextualizar a ciência de dados, assim como as competências do cientista de dados;
- Conceituar e contextualizar o ambiente *big data*;
- Identificar, a partir de estudo de literatura e análise de currículo, de que forma o profissional da informação está sendo preparado para trazer contribuições à ciência de dados.

4 A INFORMAÇÃO

Informação é um importante insumo que está presente em todos os ambientes sociais. É uma matéria-prima que perpassa desde a informação técnico-científica, aquela que percorre uma metodologia científica, indo até a mais informal, presente no cotidiano dos indivíduos. Indo além, a informação pode ser utilizada em diferentes contextos, em um cenário de imprescindível tomada de decisão, em ambiente governamental ou empresarial, tal e qual, como um fato comunicado de maneira corriqueira no dia a dia das pessoas. Desta maneira, tudo aquilo dotado de capacidade de transmitir uma ideia, um conceito, um fenômeno pode ser considerado informação, estando ela depositada em algum tipo de registro ou sendo de fonte oral.

Na ciência, a informação pode ser considerada a essência máxima do processo de produção do conhecimento científico, sendo vista como o seu sangue, posto que, sem informação a ciência não consegue se desenvolver e viver, assim como a pesquisa seria em vão e o conhecimento não existiria (LE COADIC, 1996). Isso demonstra que existe uma dinâmica entre a construção da ciência e a produção da informação de cunho técnico-científica. Uma alimentando a outra, em um processo contínuo.

O conceito de informação, nas diferentes ciências, pode abranger um conjunto significativo de definições. A concepção de informação pode ser utilizada para caracterizar uma medida de organização física; um padrão de comunicação entre um emissor e um receptor; a probabilidade de uma mensagem ser transmitida por um canal de comunicação; o significado de uma forma linguística; a redução de incertezas etc. Todavia, como defendem Capurro e Hjorland (2007), a conceituação mais relevante talvez seja aquela de informação como um objeto ou coisa (bits, por exemplo) e a informação como uma ideia subjetiva, ou seja, dependendo da interpretação de um agente cognitivo, informação como signo.

Fundamentado nessa concepção de Capurro e Hjorland (2007) de informação como objeto/coisa, torna-se pertinente citar o pensamento proposto por Buckland (1991), onde a informação passa a ser descrita por meio de três ideias principais: a informação como *processo*, como *conhecimento* e como *coisa*. A informação como processo é vista a partir da perspectiva de que quando um indivíduo é informado sobre algo, aquilo que ele conhece é modificado, nesse sentido, a informação é identificada como o ato de informar. Por sua vez, a informação como conhecimento é caracterizada por aquilo que é percebido por meio da informação como processo, o conhecimento que é comunicado, que reduz incertezas. Finalmente, a informação como coisa relaciona-se com os objetos, documentos considerados como informação, uma vez

que são relacionados como sendo informativos, dotados de conhecimento comunicável (BUCKLAND, 1991).

No trabalho aqui proposto, o conceito que melhor dialoga com os objetivos da pesquisa é o de informação como coisa, ou seja, o potencial de assimilação de informação a partir de diferentes objetos. Ademais, adota-se como princípio para o desenvolvimento da pesquisa o relacionamento intrínseco existente entre dado, informação e conhecimento (ANGELONI, 2003), estando um influenciando a construção dos demais. Assim sendo, a análise desse novo cenário de grande proliferação de dados em ambiente virtual (*big data*), a partir da disciplina CI, mostra-se como matéria dotada de potencialidades para o profissional que possui como objeto de estudo a informação.

4.1 A Ciência da Informação

Partindo da observação sobre a necessidade de compreensão dos processos no quais a informação se encontra envolvida, e sua real importância para o desenvolvimento sociocultural dos indivíduos, surge, com isso, um novo campo do saber, que tem como missão se debruçar sobre os fluxos percorridos pela informação, buscando o seu entendimento e otimização. Essa nova disciplina chama-se *ciência da informação*.

Saracevic (1996) discorre sobre a CI apontando as características gerais que constituem sua razão de existir e sua evolução, sendo a CI: uma área interdisciplinar, uma vez que dialoga diretamente com outros campos do conhecimento; está conectada de maneira profunda com a tecnologia da informação, posto que o "imperativo tecnológico" define a CI, assim como em outros ramos do saber, colaborando com o surgimento da chamada *sociedade da informação*; e, por fim, o forte vínculo que a CI possui com essa sociedade da informação, devido ao fato de ser uma disciplina com viés voltado a parte social/humanista. Portanto, para compreender o passado, o presente e o futuro desse campo, a CI, assim como os desafios enfrentados, é imprescindível entender essas três características formadoras dessa área.

Os pilares que serviram de fundação para a disciplina CI tiveram sua construção iniciada em meados da década de 1940, quando do fim da Segunda Guerra Mundial. Apesar de alguns pesquisadores argumentarem que o termo "ciência da informação" só foi proferido pela primeira vez por volta de 1960 (HEILPRIN, 1989 apud PINHEIRO; LOUREIRO, 1995), é com o desfecho dos conflitos entre as nações, resultando grande desenvolvimento científico e tecnológico, que ocorre profunda propagação de informação no seio social, a chamada *explosão*

informacional, o que, por sua vez, serve como ponto de partida para o surgimento da CI. Essa situação observada, de grande profusão de informação, mostrou-se ser, ao mesmo tempo, algo positivo, porquanto, servia de insumo no desenvolvimento cada vez maior da ciência e da tecnologia, assim como algo negativo, devido à dificuldade de recuperar essa informação produzida.

Wersig e Neveling (1975) argumentam que a CI não surgiu originada de outro campo de estudo, como a psicologia, nem como a junção entre duas outras áreas, como a bioquímica, mas a partir da necessidade de um campo de trabalho prático, chamado *documentação* ou *recuperação da informação*. Os autores ainda discorrem que, apesar dessa disciplina ter sido determinada em grande parte pelo surgimento de novas tecnologias, todavia, sua origem se encontra na intersecção entre diversas outras disciplinas, na união de uma série de interesses distintos, oriundos da ciência da computação, da biblioteconomia, filosofia e taxonomia, teoria da informação etc.

Pautada na problemática acerca da recuperação da informação, considerando-se o contexto da explosão informacional vivenciada a partir da década de 1940, surge a preocupação em estimular o debate sobre as melhores e mais adequadas soluções, garantindo-se, desta maneira, a recuperação de informações pertinentes. Cabe salientar que o termo recuperação da informação, segundo Saracevic (1996, p. 44), foi cunhado inicialmente por Mooers, em 1951, para o qual "engloba os aspectos intelectuais da descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação".

Em seu artigo, Saracevic (1996) defende justamente isso, a importância de se pensar em modelos de recuperação da informação como mecanismos fundamentais para o desenvolvimento da área de CI. O autor arrazoa que, apesar da recuperação da informação não ter sido a única responsável pelo avanço da CI, pode-se considerar como a principal motivadora de tal fato. Esse processo de pensar a recuperação da informação foi, sem dúvida, essencial para a emergência da indústria informacional.

Nessa mesma linha, Borko (1968) busca delinear uma definição que consiga abarcar a amplitude da CI. Dessa forma, a CI seria uma disciplina preocupada com o estudo das propriedades e do comportamento informacional, da natureza que demarca o fluxo informacional, assim como o processamento da informação com fim de permitir acessibilidade e otimização de uso (BORKO, 1968). Ou seja, a CI está pautada na preocupação de compreender todo o arcabouço de conhecimentos relacionados ao fluxo informacional: a

origem da informação, sua coleta, organização, armazenamento, recuperação, significação, difusão, transformação e reutilização como nova fonte informacional.

Convém ressaltar, ainda, as características de ciência social da CI, posto que, as pesquisas desenvolvidas dentro dessa área foram orientadas pela procura informacional do ser social, assim como pela necessidade de solucionar um problema social, o da informação. Le Coadic (1996) destaca que a pesquisa em CI, na busca por responder uma necessidade social, acabou prosperando em função dessa necessidade, sendo dirigida e financiada por ela. O autor também defende que, movida pela influência da Tecnologia da Informação (TI) e das máquinas de comunicar, os pesquisadores desse campo emergente tiveram como preocupação maior a utilidade, a eficácia, o prático, negligenciando, de certa forma, a teoria que fundamenta a área (LE COADIC, 1996).

Todavia, em determinado momento, a CI transcende essa característica quase exclusiva de prática de organização, tornando-se uma ciência social rigorosa, sob efeito tanto de uma crescente demanda social quanto de grandes avanços econômicos (LE COADIC, 1996). Portanto, é possível observar que a CI é forjada a partir da percepção de ausência de uma ciência destinada exclusivamente para analisar e compreender como se dá a fluidez informacional no seio da sociedade, assim como pela busca de aprimoramento do processo de recuperação da informação, numa época em que o incentivo à pesquisa técnico-científica, bem como o avanço das tecnologias, servia de catalisador para a propagação de fontes informacionais.

4.1.1 A Biblioteconomia

Verificou-se, anteriormente, que a biblioteconomia foi uma das disciplinas que colaboraram diretamente com o surgimento e progresso da CI como campo científico. Embora nos dias atuais a biblioteconomia seja uma área que está inserida dentro do amplo escopo da ciência da informação, não obstante, essa matéria tem suas origens num passado mais longínquo que a CI. Le Coadic (1996) define a biblioteconomia como sendo nem uma ciência e nem uma tecnologia rigorosa, mas, sim, uma prática voltada para a organização de bibliotecas. O autor também argumenta que devido à nova realidade vivenciada pelas bibliotecas, de dispor de acervos de diversos formatos, a biblioteca acabou por transcender o significado exato de seu nome, transformando-se, nos dias atuais, em “midiateca”.

Para rastrear o surgimento das bases e princípios que norteiam a biblioteconomia atual é necessário olhar para o passado, mais especificamente para o surgimento das primeiras

bibliotecas, onde havia compreensão da necessidade de organização do conhecimento registrado. Com a ascensão de um modelo social pautado pela organização em estados rurais artesanais, por volta de 3.000 a.C., em contraposição ao modelo de tribos pastoris nômades, vivenciado até então, inicia-se, concomitantemente, a produção documental, tanto de caráter burocrático, como testemunhal e histórico, tornando-se o ponto de partida para o surgimento do conceito de biblioteca, como defende Siqueira (2010). Mesmo ainda não havendo uma figura profissional, assim como a indistinção entre bibliotecas e arquivos, era possível perceber nesses protótipos antigos de biblioteca certa tendência de ordenação, criação de catálogos e inventários, buscando produzir mecanismos que facilitasse a organização e recuperação da informação armazenada (SIQUEIRA, 2010).

Outro ponto fundamental para a percepção do desenvolvimento da biblioteconomia enquanto disciplina está relacionado ao processo de criação e reprodução do livro como objeto armazenador de informações. Até por volta do século XV, a produção de livros era feita de forma manual, por meio dos escribas e/ou monges copistas. Ortega (2004) defende que a partir do surgimento da prensa de tipos móveis, proposta por Gutenberg em meados de 1440, e sua propagação pela Europa, por meio da dispersão de impressores, motivada pela tomada da Mogúncia, na Alemanha, em 1462, acabou ocorrendo a "vulgarização" da técnica de impressão. Esse fato contribuiu para a maior produção de livros, assim como seu barateamento e distribuição, rompendo com o monopólio que a Igreja possuía sobre a produção editorial (ORTEGA, 2004). Com esse cenário, as coleções das bibliotecas existentes, em sua maioria bibliotecas particulares ou de universidades, passam a ter acesso a um maior número de livros, impulsionando a disseminação do pensamento científico, além de contribuir com a evolução da biblioteconomia como área do saber.

Embora a maior produção de livros, observada com a inserção da prensa de Gutenberg, tenha gerado um impacto sem precedentes para a biblioteconomia, o termo em si só foi utilizado pela primeira vez em 1839, na obra *Bibliothéconomie: instructions sur l'arrangement, la conservation e l'administration des bibliothèques*, do livreiro e bibliógrafo Léopold-Auguste-Constantin Hesse. Já a sistematização das técnicas e práticas dos bibliotecários só é efetivada no século XIX (LAHARY, 1997 apud ORTEGA, 2004).

Baseando-se nesse contexto, nasce a biblioteconomia enquanto disciplina que se encarrega da organização, administração e cuidado com a gestão do livro, assim como a figura do profissional bibliotecário, preocupando-se tanto com aspectos técnicos (catalogação,

classificação, indexação), quanto com a aquisição de cultura geral (história, literatura, ciências) (SIQUEIRA, 2010).

Contudo, com o surgimento e desenvolvimento das TIs, observado na metade do século XX, o objeto de estudo da biblioteconomia sofre algumas mudanças significativas. O armazenamento informacional, que até então tinha no livro físico o meio mais difundido de conservação do conhecimento, passa a ter outros mecanismos como depósitos de informação. A evolução da computação e da internet é um dos principais fatores de mudança no paradigma biblioteconômico. Essas duas tecnologias permitiram a criação de novas formas de armazenar informação em ambiente digital, principalmente com o advento das bases de dados digitais e da computação em nuvem (SARACEVIC, 1996; LE COADIC, 1996). O ambiente da biblioteca, visto como um local de aquisição, tratamento, guarda e disseminação de objetos informacionais físicos, passa a ter que se adaptar à nova realidade de diversos tipos de mídias informacionais, inclusive aquelas disponibilizadas em meio virtual, influenciando e sendo influenciada por esse novo paradigma tecnológico.

Concomitantemente ao surgimento dessa realidade tecnológica para as bibliotecas e a biblioteconomia, surge também o pensamento de que os sistemas automatizados de recuperação informacional possam tomar o espaço das bibliotecas e dos profissionais bibliotecários, em especial as ferramentas de busca, tornando-os ultrapassados. Todavia, tal preocupação vem se mostrando completamente sem fundamento, uma vez que, a tecnologia surge com o objetivo de auxiliar as atividades humanas, buscando maximizar seu potencial. A tecnologia sozinha é insuficiente para o desenvolvimento de atividades que exigem raciocínio subjetivo. Talvez a maior preocupação que deveria ocupar o pensamento da comunidade bibliotecária seja de que forma esses profissionais podem tirar maior proveito das tecnologias, paradigmas e novas ciências que emergem a todo instante, como, por exemplo, o *big data* e a ciência de dados.

4.1.2 O Profissional da Informação

Com base no exposto, surge a indagação sobre a forma com que o chamado profissional da informação está sendo treinado para lidar com o atual cenário de extensa multiplicidade tecnológica; (apesar da expressão *profissional da informação* se relacionar de maneira geral com bibliotecários, arquivistas e museólogos, dentro do espaço da CI, neste trabalho, a ênfase será dada ao primeiro grupo). Além disso, cabe também o questionamento a respeito de quais

seriam as características exigidas desses profissionais para suprir a nova demanda existente dos usuários de informação, sejam indivíduos ou organizações.

4.1.2.1 Habilidades

Realizando um levantamento acerca de trabalhos desenvolvidos com base nas novas características que se espera dos profissionais da informação, é possível identificar algumas conclusões alcançadas pelos autores. Cunha (2000), por exemplo, levanta a questão sobre a possibilidade de diversos profissionais poderem atuar nesse ambiente de suporte informacional, como os comunicadores, cientistas da computação, cientistas da informação e gestores da informação, uma vez que, a combinação dessas áreas permite a oferta de serviços de informação gerenciados, estruturados tecnologicamente, analisados e disseminados de maneira eficaz. Cabendo a indagação de qual dessas quatro áreas conseguirá dominar esse setor de consumo de informação. Para Cunha (2000) serão aqueles que possuírem características híbridas, abarcando um pouco de cada disciplina. A autora segue afirmando que a "atividade de informação é muito vasta, envolve muitos aspectos para que seja coberta por um único profissional com uma formação única; esta abertura e esta troca com profissionais de várias áreas proporciona [...] possibilidades de um trabalho mais diversificado e mais rico" (CUNHA, 2000, p. 3).

Por sua vez, Targino (2000) buscou sintetizar os requisitos básicos que um profissional que lida com informação necessita possuir. Esses requisitos vão desde ter *visão gerencial* – o que permite a esses profissionais tomar decisões de maneira racional e eficiente, como questões relacionadas ao custo da informação e o seu caráter estratégico –; a *capacidade de análise* – servindo como aporte no momento da tomada de decisão, face aos diversos tipos de suportes, a variedade de uso da informação e as distintas demandas informacionais –; a *criatividade* – característica que permite agir de forma original diante de situações atuais, permitindo buscar novas soluções para problemas antigos –; e, finalmente, a *atualização* – processo que está diretamente ligado à educação continuada, onde espera-se do profissional da informação a constante busca de novas tecnologias e técnicas para auxiliar o usuário/cliente no momento da busca informacional.

Valentim (2000), em seu trabalho a respeito das competências que caracterizam o profissional da informação tido como "moderno", esboça 4 conjuntos de habilidades consideradas imprescindíveis, a saber: competências de *comunicação*, *técnico-científicas*, *gerenciais* e *sociais e políticas*. As competências de comunicação dizem respeito aos produtos

que possibilitam o processo de comunicação entre o usuário e a fonte informacional (bibliografias, catálogos, índices etc.). Por sua vez, as competências técnicos-científicas são aquelas associadas ao desenvolvimento e execução de atividades relacionadas com o tratamento de fontes de informação, nos diferentes suportes, unidades e serviços de informação. Seguindo, as competências gerenciais abordam as atividades de formulação, administração, organização e coordenação de unidades, sistemas, projetos e serviços de informação. Em conclusão, as competências sociais e políticas estão relacionadas com as ações do profissional da informação no âmbito da sociedade, buscando viabilizar seu desenvolvimento pessoal, institucional e social (VALENTIM, 2000).

Já Ferreira (2003) apresenta um conjunto de habilidades consideradas como essenciais pelas organizações na prática de gestão do conhecimento. Segundo o autor, essas funções estão relacionadas com a execução de atividades na área de classificação das fontes informacionais, acesso, recuperação e análise da informação, desenvolvimento de produtos e serviços a partir da informação, união do conhecimento com a experiência das pessoas dentro da organização e trabalhar com a proteção do conhecimento (FERREIRA, 2003, p. 46). Além disso, a partir da confecção de um *ranking* de habilidades mais demandadas pelo mercado, o autor consegue identificar 15 habilidades, sendo que as cinco primeiras são: conhecer o ambiente de negócios da informação; ter facilidade de trabalhar em grupo; ter discernimento sobre informações relevantes e a relevância das informações; possuir capacidade de utilizar equipamentos eletrônicos e operação de *softwares* específicos; e conhecimento sobre bases de dados (FERREIRA, 2003). Nota-se, a partir da análise do autor, a busca por um profissional que detenha conhecimentos relacionados com gestão de pessoas, liderança e ambiente organizacional (administração), assim como princípios voltados para a busca e análise de fontes informacionais (CI), juntamente com habilidades relativas a *hardware* e *software* (ciência da computação).

Por seu turno, o Incite (2005 apud BELLUZZO, 2011, p. 64-65) aborda cinco grupos de habilidades consideradas fundamentais para um profissional da informação (mais especificamente, um profissional bibliotecário). O grupo *Informação* (I) são as competências que todos profissionais devem possuir, em maior ou menor nível (as habilidades essenciais). O grupo *Tecnologias* (T) vai além das competências essenciais, e diz respeito a utilização de instrumentos mediados pelas tecnologias emergentes. Por sua vez, o grupo *Comunicação* (C) aborda a interdependência entre as noções de informação e comunicação, assim como a complementaridade de seus meios. Já o grupo *Gestão* (M) relaciona-se com a necessidade da

gestão da informação, bem como a repercussão que há sobre a qualidade da informação que é gerida. Por fim, o grupo *Outros Saberes* (S) é aquele que remete aos conhecimentos que permitem ao profissional da informação ser versado em áreas conexas. Desta forma, percebe-se, mais uma vez, a exigência que a CI e, consequentemente, o profissional da informação possua conhecimento holístico sobre campos relevantes para seu desenvolvimento, confirmando seu caráter interdisciplinar.

Santa Anna, Pereira e Campos (2014) utilizam a terminologia cunhada pela Federação Internacional de Documentação e Informação (FID) para se referirem a esses novos profissionais, os Modernos Profissionais da Informação (MIP). Além disso, os autores também retomam as quatro habilidades propostas por Valentim (2000), defendendo, contudo, que novas competências necessitam ser incorporadas, buscando inserir "novas competências de cunho tecnológico, educacional e cultural" (SANTA ANNA; PEREIRA; CAMPOS, 2014, p. 81).

Por fim, analisando as competências pessoais esperadas de um profissional da informação, expostas na Classificação Brasileira de Ocupações do Ministério do Trabalho[1], é possível identificar competências relacionadas com os seguintes tópicos: atualização constante, liderança, trabalho em equipe, capacidade de análise e síntese, conhecimento de outros idiomas, capacidade de comunicação, senso de organização, capacidade empreendedora, capacidade de concentração, proatividade, criatividade etc.

As tecnologias mostram-se como as principais variáveis modificadoras do ambiente de atuação do profissional da informação. Coneglian, Gonçalvez e Santarém Segundo (2017) afirmam que cabe ao profissional, que tem a informação como insumo de trabalho, possuir conhecimento e domínio no uso da TI. Além disso, desses profissionais espera-se o desenvolvimento de produtos de informação, visando o uso interno e externo de suas organizações, como por exemplo: a criação de bases de dados, páginas virtuais, arquivo de texto etc. Já entre as competências que se espera desses profissionais da informação, pode-se elencar "visão globalizada; buscar desafios; investir em novas oportunidades; comunicar-se com eficácia; criar parcerias e alianças e construir um ambiente tendo como base o respeito e a confiança" (CONEGLIAN; GONÇALVEZ; SANTARÉM SEGUNDO, 2017, p. 132).

A partir do cenário apresentado, verifica-se que o contexto de atuação do profissional da informação vem mudando incessantemente, assim como as atribuições que se espera deles. O fator que mais corrobora com essa mudança vem sendo o surgimento de novas tecnologias

[1] Disponível em: <http://www.mtecbo.gov.br/cbosite/pages/pesquisas/ResultadoFamiliaAtividades.jsf>. Acesso em: 11 jul. 2018.

de comunicação e informação. Desta maneira, o mercado passa a buscar profissionais que estejam em constante processo de aprendizagem, evoluindo profissionalmente junto com o avanço das novas tecnologias e modelos. Um dos desafios surgido, recentemente, diante do profissional da informação é o *big data*, que nas palavras de Coneglian, Gonçalvez e Santarém Segundo (2017, p. 132) "é caracterizado por volumes de dados extremamente densos e que necessitam de competências, habilidades e ferramentas para que essa informação possa ser encontrada; para que isso seja possível, ela necessita ser tratada, analisada e disponibilizada em tempo hábil", podendo o profissional da informação desempenhar papel fundamental no decorrer desse processo.

5 BIG DATA

Atualmente vive-se um fenômeno de produção em larga escala de dados. Tendo como parâmetro o ano de 2012, por volta de 2,5 exabytes de dados foram criados diariamente, e esse número segue dobrando a cada 40 meses. Mais dados cruzam a internet a cada segundo do que o que foi armazenado em toda a internet há apenas 20 anos. A título de exemplo, estima-se que o Walmart coleta mais de 2,5 petabytes de dados a cada hora de suas transações com clientes, sendo que um petabyte equivale a um quatrilhão de bytes, e um exabyte é 1.000 vezes esse valor, ou um bilhão de gigabytes (MCAFEE; BRYNJOLFSSON, 2012, p. 62).

Esse fenômeno descrito é motivado, principalmente, pela "drástica redução de preços para o armazenamento das informações; a explosão de aplicações disponíveis na internet (*e-commerce*); a popularização de sensores conectados – internet das coisas, pesquisas científicas – ao projeto genoma; e, as redes sociais" (VICTORINO et al., 2017, p. 230). Com o surgimento desse cenário, surge também a necessidade de se pensar em soluções que possibilitem melhorar o tratamento e uso dos dados que são produzidos, objetivando beneficiar a tomada de decisões.

O termo *big data* emerge, então, como um modelo de representação das características observadas no contexto de grande profusão de dados. A partir da inserção dos computadores no seio social, meio século atrás, os dados começam a ser acumulados, permitindo o surgimento de algo novo. O mundo passa a não estar apenas repleto de informação, mas a informação começa a ser acumulada com mais rapidez. Desta forma, o *big data* surge na esteira do avanço de ciências como a astronomia e a genômica, embora, o termo, atualmente, esteja migrando para as mais diversas áreas do conhecimento (MAYER-SCHONBERGER; CUKIER, 2013).

Davenport (2014) afirma que *big data* nada mais é que um conjunto de dados grande o suficiente para não caber em repositórios usuais, ou seja, dados volumosos demais para serem guardados em servidores comuns. Além disso, ainda segundo o autor, esses dados não são estruturados o suficiente para serem alocados em bancos de dados tradicionais – organizados em linhas e colunas –, ou fluídos demais para serem acomodados em estruturas estáticas de armazenagem.

Não obstante, a definição mais difundida entre os estudiosos da área advém dos chamados "3 Vs". Laney (2001) é o primeiro a analisar o fenômeno de grande produção de dados à luz de seu *volume*, *velocidade* e *variedade* (3 Vs). O autor observa que, com o avanço do comércio eletrônico, o que se vê é uma produção de dados em escala cada vez maior (volume) – corroborado pelo progresso crescente da capacidade de armazenamento dos bancos

de dados –, com elevada rapidez de produção (velocidade) – em grande parte devido ao constante avanço das tecnologias de processamento –, assim como o diversificado conjunto de formatos de dados que estão disponíveis (variedade) – textos, imagens, vídeos etc.

Com o passar do tempo, no entanto, novos trabalhos acerca da temática *big data* foram surgindo, e, consequentemente, novas proposições foram sendo firmadas a partir dos 3 Vs iniciais de Laney. Gandomi e Haider (2015, p. 139), por exemplo, acrescentam mais três dimensões: *veracidade*, *variabilidade* e *valor*. Segundo os autores, a veracidade está relacionada com a insegurança inerente que algumas fontes de dados possuem – ou seja, lidar com dados imprecisos e incertos pode ser outra faceta do *big data* –; por sua vez, a variabilidade está relacionada com a variação nas taxas de fluxo dos dados – esta dimensão relaciona-se com a velocidade de produção de dados, podendo haver alta e baixa velocidade (variabilidade) –; por último, a característica valor é considerada um atributo definidor do *big data* – os dados recebidos em sua forma original geralmente possuem baixo valor em relação ao seu volume, podendo ser gerado alto valor a partir da análise de grandes volumes desses dados.

Embora possam existir diversas dimensões vinculadas ao *big data*, de acordo com a visão específica de cada autor, é importante destacar que cada dimensão não é independente em relação às demais. A partir do momento que uma característica muda, existe grande probabilidade de isso se refletir nas outras características (GANDOMI; HAIDER, 2015), demonstrando forte ligação entre as dimensões relacionadas ao ambiente *big data*.

Ainda no tocante às definições existente sobre a matéria *big data*, Rodrigues, Nóbrega e Dias (2017, p. [5]), na busca de maior entendimento sobre o assunto, elaboram um quadro em que trazem definições e contextos que estão relacionados com o ambiente em questão, auxiliando, assim, "na compreensão do fenômeno e das concepções que ele adquire em campos distintos como Computação, Economia, Ciência da Informação".

Quadro 1 – Definições e contextos de big data

Autor	Definição
Di Martino et al. (2010, p. 5)	"É um campo emergente onde inovadora tecnologia oferece alternativas para resolver os problemas inerentes que surgem quando se trabalha com grandes quantidades de dados, fornecendo novas maneiras de reutilizar e extrair valor a partir de informação".
Manika et al. (2011, p. 1)	"Refere-se a um banco de dados cujo tamanho vai além da capacidade do software de banco de dados e ferramentas típicas para capturar, armazenar, gerenciar e analisar".

Boyd e Crawford (2012, p. 663)	"Um fenômeno cultural, tecnológico, acadêmico e que repousa sobre a interação de tecnologia, análise e mitologia".
Dumbill (2012, *online*)	"São dados que excedem a capacidade de processamento dos sistemas de banco de dados convencionais".
Mayer-Schonberger e Cukier (2013, p. 4)	"Refere a trabalhos em grande escala que não podem ser feitos em escala menor, para extrair novas ideias e criar novas formas de valor de maneira que alterem os mercados, as organizações, a relação entre cidadãos e governos, etc.".
Moura e Amorim (2015, p. 2)	"Expõe uma nova geração de tecnologia e arquitetura, destinada a extrair valor de uma imensa variedade de dados permitindo alta velocidade de captura, descoberta e análise, transformando dados em informações valiosas".
Goularte, Zilber e Pedron (2015, p. 3)	"Não se trata apenas de uma ferramenta, mas é, em verdade, uma geração de novas tecnologias e arquiteturas projetadas para extrair valor econômico de grandes volumes de dados".
Menezes, Freitas e Parpinelli (2016, p. 1)	"Inúmeras bases de dados estão tendenciando a possuir grande volume, alta velocidade de crescimento e grande variedade. Esse fenômeno é conhecido como *Big Data* e corresponde a novos desafios para tecnologias clássicas como Sistema de Gestão de Banco de Dados Relacional".

Fonte: Rodrigues, Nóbrega e Dias (2017, p. [5-6]).

A partir das definições apresentadas no Quadro 1, alguns pensamentos pertinentes acerca do *big data* podem ser extraídos. Muitos autores destacam como sendo o objetivo maior da análise de grandes volumes de dados a geração de "informações valiosas" e, consequentemente, "novas ideias" que irão auxiliar as organizações no momento da tomada de decisão. Outros evidenciam os desafios oriundos dessa nova realidade de proliferação de dados, refletindo, assim, na superação da "capacidade de processamento" das tecnologias tradicionais. O que, por sua vez, irá repercutir no desenvolvimento de "uma geração de novas tecnologias e arquiteturas", destinadas a otimizar o processo de "captura, descoberta e análise" desse grande volume de dados. A seguir serão descritos os mais importantes princípios e tecnologias criadas para contribuir com o tratamento e análise do ambiente *big data*.

5.1 Princípios e Tecnologias em Big Data

Os problemas gerados para analisar a enorme quantidade de dados, exposta anteriormente, podem se apresentar de várias maneiras. Em certo momento, as técnicas tradicionais usadas para trabalhar com dados não conseguem mais acompanhar o ritmo de produção desses dados, disponíveis em diversos formatos.

Todavia, cabe destacar que os problemas que o *big data* traz consigo não são uma percepção recente, embora tenha ganhado mais espaço para discussão nas últimas décadas. O problema da "armazenagem" e "compreensão" de grandes quantidades de dados já era identificado na década de 1960, quando a empresa norte-americana RAND trabalhava em um projeto de *Relational Data File* (Arquivo de Dados Relacionais) – sistema projetado para analisar de maneira lógica uma grande coleção de dados factuais. Desta forma, em 1967, dois cientistas da computação encontravam dificuldades em trabalhar com grandes conjuntos de dados, pois notaram que, junto com o vasto volume de dados, vinha também uma variedade de problemas de caráter lógico e linguístico, de *hardware* e *software*, práticos e teóricos, trazendo prejuízo aos seus empreendimentos (CRAWFORD; MILTNER; GRAY, 2014, p. 1.664).

Assim sendo, como discorrem Victorino et al. (2017), várias pesquisas despontam motivadas pela busca em desenvolver novas tecnologias aptas a lidar com os problemas de armazenamento e processamento desse vasto volume de dados, produzidos em grande velocidade e de forma variada. A seguir são apresentados alguns princípios fundamentais dentro do ambiente *big data*, assim como tecnologias úteis para facilitar a ação de análise e interpretação desses dados. Os tópicos escolhidos para breve descrição se baseiam em Davenport (2014, p. 112) e Victorino et al. (2017).

5.1.1 Analytics em Big Data

Para falar sobre o uso de *analytics* no ambiente *big data* é preciso primeiro identificar as raízes e conceituar o termo. A definição de *analytics* está diretamente associada com a expressão *business intelligence* (BI), que, por sua vez, surge por volta da década de 1950 (DAVENPORT, 2014), quando pesquisadores em inteligência artificial passam a utilizá-la. Todavia, o BI tornou-se um termo popular nas comunidades de negócios e de TI apenas nos anos de 1990. No final dos anos 2000, o conceito de *business analytics* foi introduzido para

representar o componente analítico no BI. Mais recentemente, o termo *big data analytics* é usado para descrever os conjuntos de dados e técnicas analíticas em aplicações que são tão grandes (de terabytes a exabytes) e complexas (de dados vindos de sensores a mídias sociais), que exigem avançadas e exclusivas tecnologias de armazenamento, gerenciamento, análise e visualização de dados (CHEN; CHIANG; STOREY, 2012, p. 1.166).

Desta forma, embora a utilização do termo *analytics* em BI tenha se difundido com maior força na atualidade, suas raízes datam de muito antes. O que se vê hoje é uma adaptação, para o ambiente *big data*, daquilo que já era feito com dados comuns. Davenport (2014, p. 4) destaca as diferenças básicas entre o *analytics* tradicional, utilizado junto ao *business intelligence*, e o *analytics* utilizado em grandes volumes de dados. No primeiro, segundo o autor, os dados são formatados em linhas e colunas – podendo ser armazenados em bancos de dados convencionais –; o volume dos dados está na casa dos terabytes ou menos; o fluxo de dados é *pool* estático; os métodos de análises são baseados em hipóteses; e o objetivo principal é dar suporte ao processo decisório da organização. Já o *analytics* em *big data* tem seus dados em formatos não estruturados – o que exige bancos de dados especiais para armazená-los –; o volume dos dados gira em torno de 100 terabytes a petabytes; o fluxo de dados é constante; o método de análise é por meio de *machine learning* (aprendizado de máquina); e o objetivo principal é gerar produtos baseados em dados.

Logo, o que se vê é um princípio que já era utilizado anteriormente, só que desta vez num contexto distinto. O *analytics* em *big data*, chamado por Davenport (2014) de *analytics* 3.0, está associado a grandes volumes de dados. Esses conjuntos de dados, ao contrário do *analytics* tradicional, exigem novos dispositivos tecnológicos capazes de resistir ao fluxo constante de dados, em sua mais variada forma.

5.1.2 Data Mining e Machine Learning

Com a ascensão do poder de armazenamento dos *hardwares* existentes, inicia-se, em conjunto, o processo de guarda de dados em escalas cada vez maiores. Todavia, juntamente com essa nova realidade, também emerge a indagação do que fazer com esse grande volume de dados armazenados. Surge o conceito de *data mining* (mineração de dados). Para Han, Kamber e Pei (2012, p. 8), o *data mining* é o processo de descoberta de padrões e conhecimentos, considerados pertinentes, a partir de grandes quantidades de dados. Para os autores, as fontes

de dados podem ser diversas: bases de dados, *data warehouses*[2], internet, outros repositórios de informações ou dados transmitidos para o sistema de forma dinâmica. O processo de mineração dos dados é feito por *softwares* baseados em técnicas estatísticas e matemáticas, em que se busca extrair os dados considerados significativos a partir do grande volume de dados originário.

Concomitantemente, em algumas circunstâncias, é possível que "máquinas", isto é, computadores e algoritmos, consigam aprender através de dados, processo conhecido como *machine learning*. Certamente existem alguns casos de sucesso envolvendo o "aprendizado de máquina", como análises de texto, reconhecimento de imagem e escrita feita a mão, filtros para *spam* e detecção de fraudes no cartão de crédito (FRICKÉ, 2015). Nesse contexto, o aprendizado de máquina, segundo Davenport (2014, p. 112), consiste em um software desenvolvido especificamente para conseguir identificar o melhor modelo para análise de determinado conjunto de dados.

De acordo com Han, Kamber e Pei (2012, p. 26), é possível encontrar similaridades entre o aprendizado de máquina e a mineração de dados. Em se tratando de tarefas de classificação e agrupamento (*clustering*), as pesquisas sobre aprendizado de máquina frequentemente focam na precisão do modelo. Por sua vez, somando-se à precisão, as pesquisas em mineração de dados colocam forte ênfase na eficiência e escalabilidade dos métodos de mineração em grandes conjuntos de dados, assim como sobre formas de lidar com tipos complexos de dados e explorar novos métodos alternativos (HAN; KAMBER; PEI, 2012).

5.1.3 Hadoop

A ferramenta Haddop, segundo Victorino et al. (2017, p. 231), "é uma abordagem que busca dividir os problemas complexos do *Big Data* em pequenas unidades de trabalho e processá-las em paralelo". Por sua vez, Davenport (2014, p. 113) discorre sobre o Hadoop afirmando ser "um conjunto de ferramentas de *software* de código aberto e um *framework* para distribuir os dados em vários computadores; trata-se de um ambiente unificado de armazenamento e processamento altamente escalonável para grandes e complexos volumes de dados". O *software* costuma ser chamado de Apache Hadoop, uma vez que sua versão mais conhecida foi desenvolvida pela organização sem fins lucrativos *Apache Software Foundation*

[2] Estrutura "que contém uma grande variedade de conjuntos de dados com uma grande variedade de finalidades e temas" (DAVENPORT, 2014, p. 114).

(ASF). No entanto, como o programa foi desenvolvido em código aberto, surgiram várias outras versões dele: *Cloudera Hadoop, Hortonworks Hadoop, EMC Hadoop, Intel Hadoop, Microsoft Hadoop* etc. (DAVENPORT, 2014).

O Hadoop vem sendo fundamental para tornar ágil o processo de análise de dados. Em se tratando de desenvolvimento de *software*, as chamadas "práticas ágeis" estão associadas a ciclos rápidos de produção, interação mais próxima entre desenvolvedores e consumidores e a efetuação de testes. A análise de dados tradicional tem sido dificultada por persistentes processos de *turn-around* – se um cálculo é iniciado, pode levar horas para ser finalizado, talvez até dias. Mas o Hadoop, de acordo com Loukides (2010, p. 6), facilitou a construção de *clusters*[3] capazes de executar cálculos em grandes conjuntos de dados de maneira rápida. O *software*, ainda segundo Loukides (2010, p. 6), consegue processar dados assim que eles chegam, fornecendo resultados intermediários quase em tempo real. Esta análise quase em tempo real é o que permite recursos como o *trending topics* em redes sociais como o Twitter.

5.1.4 MapReduce

MapReduce, conforme Davenport (2014, p. 112), é "um *framework* arquitetônico no qual o Hadoop se baseia". Esse *framework*[4], ainda de acordo com o autor, foi desenvolvido pela empresa Google com o objetivo de distribuir o processamento de dados através de um nó de computadores ligados entre si. Corroborando com Davenport, Loukides (2010, p. 5) afirma que o MapReduce possui como lógica básica a estratégia "dividir para conquistar", distribuindo um grande problema através de um grande grupo de computadores interligados.

Segundo Dean e Sanjay (2008 apud VICTORINO et al., 2017, p. 231), a lógica de atuação do MapReduce pode ser dividida em duas etapas: o *mapeamento* e a *redução*. O mapeamento consiste numa divisão, feita por um nó mestre, em diversos subconjuntos menores; então, através do controle de um rastreador de trabalho, um nó trabalhador fica responsável por processar os dados de um subconjunto menor e armazenar os resultados no sistema de arquivos local, possibilitando o acesso de um redutor. Por sua vez, a redução "analisa e reúne os dados de entrada a partir das etapas de mapeamento. Pode haver múltiplas tarefas de redução para paralelizar o processamento. Estas são executadas nos nós trabalhadores sob o controle do rastreador de trabalho" (DEAN; SANJAY, 2008 apud VICTORINO et al., 2017, p. 231).

[3] Agrupamento de computadores ligados entre si que trabalham em conjunto, proporcionando melhor desempenho para o sistema.
[4] Abstração da qual derivam vários projetos de software, uma vez que possuem em comum similaridades de código.

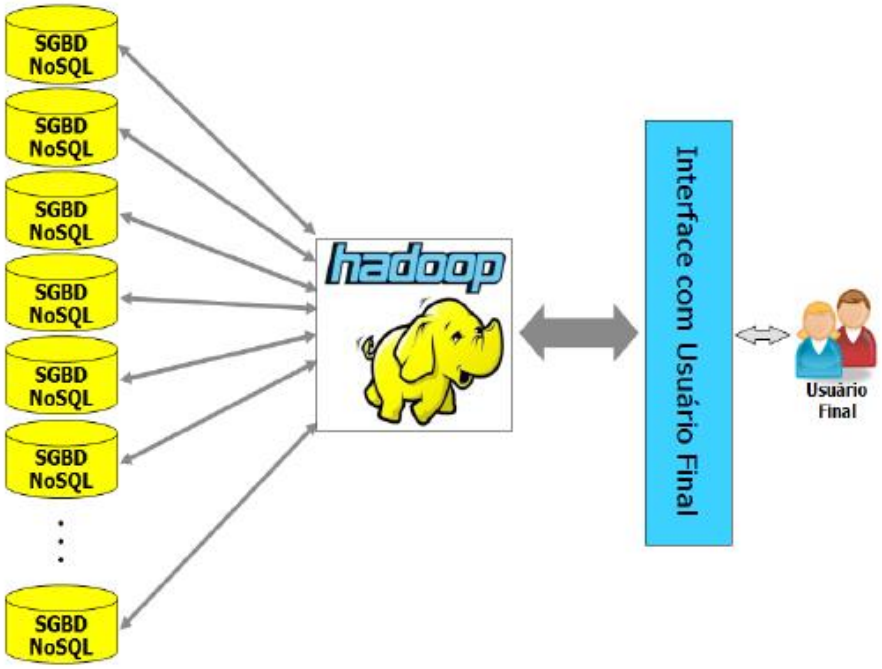
5.1.5 NoSQL

Mencionou-se, até agora, algumas tecnologias e princípios que são utilizados na otimização do processamento de grandes quantidades de dados, visando extrair dados significativos. No entanto, o entendimento sobre bancos de dados é essencial para todos que almejam trabalhar com grandes conjuntos de dados. O ambiente *big data*, como já dito anteriormente, possui como uma de suas características a variedade de formatos de seus dados. Isso significa que são gerados tanto dados estruturados – aqueles que podem ser armazenados em bancos de dados convencionais (linha e coluna) –, quanto dados não estruturados – dados que não preenchem os requisitos para armazenagem em bancos de dados relacionais. Tendo isso em mente, surge uma nova geração de bancos de dados, os chamados "bancos de dados NoSQL", ou *Non-Relational databases*. Eles foram projetados para serem distribuídos em vários nós, provendo "consistência eventual", mas não consistência absoluta, e possuindo um esquema bem flexível (LOUKIDES, 2010, p. 5).

Seguindo nessa mesma linha, Indrawan-Santiago (2012 apud VICTORINO et al., 2017, p. 233) assevera que "os bancos de dados NoSQL foram concebidos para dar suporte a processos operativos que manipulam um grande volume de dados em diversos formatos em um tempo aceitável, e não para fazer análises mais elaboradas, a fim de proporcionar suporte à decisão". Ainda segundo a autora, entre as principais vantagens dos bancos de dados NoSQL estão: ser flexível em suas estruturas, alta escalabilidade horizontal, oferecer suporte a dados não estruturados e o processamento distribuído.

No desenvolvimento de um sistema de *big data*, tanto o Hadoop quanto o NoSQL são considerados recursos fundamentais para sua implantação. Victorino et al. (2017, p. 232) arrazoam que numa arquitetura básica para *big data*, os dados são distribuídos em diversos bancos de dados NoSQL; posteriormente, o Hadoop, funcionando como um *middleware*, mapeia os dados distribuídos e aplica as transformações necessárias (MapReduce), possibilitando a consolidação dos resultados e a geração de uma resposta que é apresentada ao usuário do sistema.

Figura 1 – Arquitetura padrão de um ambiente big data



Fonte: Victorino et al. (2017, p. 232) adaptado de Davenport (2014).

A título de finalização, é importante destacar que os aspectos sobre as tecnologias e princípios utilizados em ambientes de *big data*, aqui apresentados, são de caráter ilustrativos, não sendo o objetivo deste trabalho aprofundar em seus fundamentos e conceitos. Além disso, cabe ressaltar que as tecnologias apresentadas aqui não possuem natureza exaustiva, uma vez que, optou-se por fazer um apanhado geral das ferramentas e princípios considerados essenciais no desenvolvimento de um sistema focado em *big data*. Ou seja, conceitos básicos que precisam ser conhecidos pelos profissionais que almejam trabalhar com dados em grandes volumes – os cientistas de dados.

5.2 Qualidade de Dados em Big Data

Desde o surgimento do fenômeno da explosão informacional, a relevância da informação, relacionando-se com o contexto da tomada de decisão e do desenvolvimento científico, passa a ter maior destaque. Não obstante, surge o debate a respeito da qualidade dessa informação que é produzida e quais seriam os parâmetros necessários para avaliar a qualidade de determinado conjunto de dados e informações.

Esse debate sobre a qualidade das informações dentro da CI ganha força a partir do seminário promovido pela *Nordic Council for Scientific Information and Research Libraries*, ocorrido no ano de 1989. O encontro foi visto como um importante esforço na busca da teorização sobre o assunto e do desenvolvimento de critérios e atributos que pudessem trazer maior qualificação para os dados e informações (FAGUNDES; MACEDO; FREUND, 2018, p. 197).

A partir disso, alguns autores começam a propor pesquisas empíricas tentando identificar as dimensões que conseguiriam trazer maior entendimento à qualidade dos dados. Wang e Strong (1996 apud FAGUNDES; MACEDO; FREUND, 2018), por exemplo, enxergavam o conceito de qualidade de dados como sendo multidimensional. Os autores então propuseram um quadro conceitual em que a qualidade dos dados era vista a partir de quatro aspectos: a acessibilidade dos dados aos usuários; a facilidade de compreensão da sintaxe e semântica dos dados; a utilidade dos dados; e a credibilidade dos dados para os usuários. Dando andamento aos estudos, os autores conseguiram definir quatro grupos de categorias, incluindo o total de 15 dimensões (atributos): *Intrínseca* (precisão, objetividade, credibilidade e fidedignidade); *Contextual* (relevância, valor agregado, atualização, completeza e valor apropriado); *Representacional* (interpretável, fácil de entender, representação concisa e representação consistente); e *Acessibilidade* (acessível e seguro).

Por sua vez, o surgimento do novo fenômeno de maior produção, processamento e variedade de dados em ambiente virtual (*big data*) traz consigo novas abordagens e procedimentos para a geração, seleção e manipulação dos dados (FAGUNDES; MACEDO; DUTRA, 2017), o que, por conseguinte, influencia as discussões relacionadas com a temática qualidade de dados, que passa a ser tratada como qualidade de dados em *big data*.

Fagundes, Macedo e Dutra (2017), por exemplo, buscam criar um paralelo entre os aspectos utilizados na avaliação de qualidade das informações e as dimensões que caracterizam o ambiente *big data*. Os autores utilizam o critério de qualidade das informações denominado *Methodology for Information Quality Assessment* (AIMQ), que também utiliza 15 critérios para definir essa qualidade, sendo: acessibilidade, suficiência, credibilidade, completeza, representação concisa, representação consistente, facilidade de operação, exatidão, interpretabilidade, objetividade, relevância, reputação, segurança, atualidade e compreensibilidade. Já os aspectos selecionados para representar o ambiente *big data* foram: volume, velocidade, variedade, valor, veracidade, variabilidade e visualização. A partir da análise feita, os autores conseguiram identificar a existência de relações entre todos os critérios

de qualidade da informação, propostos pela metodologia AIMQ, e os sete V's usados na representação do *big data*. Não obstante, não foi possível propor um modelo de qualidade informacional eficiente apenas com os critérios utilizados em sua análise (FAGUNDES; MACEDO; DUTRA, 2017, p. [14]).

Também tratando sobre a qualidade dos dados e informações em ambientes de *big data*, Firmani et al. (2016) abordam a dificuldade de propor uma definição única sobre qualidade dos dados no *big data*. Desta forma, defendem os autores, existem várias noções sobre qualidade, aplicadas nos diferentes tipos de dados, que devem ser cuidadosamente consideradas quando se lida com grandes volumes de dados e suas análises.

Assim como nos estudos anteriores, Firmani et al. (2016) discorrem sobre um conjunto de dimensões capazes de capturar aspectos importantes da qualidade dos dados e informações, sendo que essas dimensões podem ser divididas em oito conjuntos: acurácia, completude, consistência, redundância, legibilidade, acessibilidade, confiança e utilidade. Nota-se grande semelhança entre os modelos propostos por Wang e Strong (1996 apud FAGUNDES; MACEDO; FREUND, 2018), Fagundes, Macedo e Dutra (2017) e Firmani et al. (2016), indicando que provavelmente eles possuem uma origem comum.

Apoiando-se nas pesquisas realizadas sobre o tema qualidade dos dados em ambientes *big data* é possível perceber a complexidade existente para criar modelos de análise que sirvam para todos os tipos de cenários de dados. Quando a proposta é direcionada a um contexto comum de dados as dificuldades já se mostram reais, quando o foco é direcionado para o *big data* os obstáculos podem ser ainda maiores, devido a suas características singulares. Por exemplo, em comparação com um cenário tradicional de dados, o ambiente *big data* traz consigo duas complexidades adicionais: possui variadas fontes para os dados (fontes de origem humana, fontes mediadas por processos e fontes geradas por máquinas) e ser altamente desestruturado e desprovido de esquemas (FIRMANI et al., 2016). O debate a respeito do tema pode ser desenvolvido a partir de várias áreas, inclusive de campos que surgem para tratar de forma específica a problemática *big data*, como a ciência de dados, assunto da próxima seção.

6 CIÊNCIA DE DADOS

Como exposto anteriormente, o momento atual é de um cenário de vasta produção de dados, em grande parte causada pelo maior desenvolvimento das tecnologias de computadores, redes e sensores, bem como, do barateamento e rapidez do processamento e armazenamento que é feito pelas máquinas. Em 2013, por exemplo, de acordo com a *International Data Corporation* (EMC/IDC), existia cerca de 4 zetabytes de informações armazenadas no mundo, e este montante vem dobrando a cada dois anos. Para se ter uma ideia, 1 zetabyte corresponde à 10^{20} bytes, o que equivaleria ao poder de armazenamento de centenas de CD-ROMs distribuídos para cada ser humano existente na Terra (FRICKÉ, 2015).

Diante desse contexto, acaba emergindo a necessidade de um profissional que consiga apresentar soluções e extrair valor dessa grande quantidade de dados, o que, por sua vez, colabora com o surgimento de um novo campo de estudo, a chamada *ciência de dado*s. Essa matéria manifesta-se com uma conjuntura altamente interdisciplinar, baseando-se em técnicas e teorias advindas de diversos campos das ciências básicas. Com esse viés interdisciplinar, a ciência de dados aflora como um importante elemento dentro de vários setores, tais como saúde, indústria, finanças, segurança pública, mobilidade urbana, esportes, comunicação etc. (PORTO; ZIVIANI, 2014).

Davenport e Patil (2012) afirmam que esta é a carreira mais "sexy" do século 21. Tal afirmação surge oriunda, em grande parte, pela demanda existente no mercado por profissionais que consigam lidar com o cenário *big data* e sua consequente necessidade de obtenção de resultados a partir desses dados, gerando vantagens competitivas para as organizações. Corroborando com esse contexto de procura por especialistas em ciência de dados, uma análise do *Google Trends*, em 2012, mostrou haver uma crescente busca, de usuários de diversos países, por informações relacionadas com os termos "data scientist" e "data science", buscas quase sempre combinadas com termos sobre formação profissional, cursos, salário, habilidades necessárias e certificação profissional (CURTY; SERAFIM, 2016). Desta forma, percebe-se haver uma compreensão crescente, tanto das organizações quanto dos profissionais, acerca das potencialidades e oportunidades que surgem em conjunto com essa nova área.

Quanto ao nascimento da disciplina, apesar do termo *ciência de dados* ser relativamente novo[5], ao menos em se tratando da delimitação do escopo de atuação de um profissional da área

[5] Finzer (2013) atribui a aparição do termo "ciência de dados" ao autor William S. Cleveland, concebido em 2001, na obra *Data science: an action plan for expanding the technical areas of the field of statistics*. Todavia, Press (2013), em sua cronologia sobre o desenvolvimento da disciplina ciência de dados, discorre sobre o aparecimento

e das habilidades necessárias para isso – assim como da oferta de cursos relacionados especificamente ao assunto –; todavia, a busca de compreensão de dados através do trabalho de cientistas, estatísticos, profissionais da informação e cientistas da computação já encontrava espaço de discussão há muito tempo (PRESS, 2013).

Em 1962, John W. Tukey, um estatístico norte-americano, já defendia a necessidade da realização de análises a partir de dados. O autor de *The future of data analysis* explicava que durante muito tempo ele acreditou estar interessado apenas nas inferências feitas do particular para o geral, proporcionadas pelos métodos oriundos da estatística clássica. Mas, a partir de determinado momento, percebeu que o seu interesse estava de fato na área de análise de dados (PRESS, 2013). Aliás, ele possuía a ideia de que partes da estatística deveriam passar a fazer a análise de dados, assumindo, assim, características de ciência, ao invés de ser vista apenas como um ramo da matemática.

Por meio de uma linha temporal, traçada por Press (2013), é possível perceber como a ciência de dados evoluiu ao longo das últimas décadas (mesmo que o uso do termo só se torne mais comum a partir da década de 1990/2000). Fundamentado nessa cronologia, nota-se que os primeiros princípios norteadores da atual ciência de dados começam a se desenvolver na década de 1960, inicialmente com os trabalhos de Tukey. Desde então, o que se vê é o surgimento de livros, publicações seriadas, artigos, *workshops*, entidades e encontros de especialistas que abordam, inicialmente, os temas processamento e análise de dados e, posteriormente, *big data* e ciência de dados, sendo alguns exemplos: o livro *Exploratory data analysis*, em 1977, de John W. Tukey; o livro *From data mining to knowledge discovery in databases*, em 1996, de Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth; o periódico *The Journal Data Mining and Knowledge Discovery*, lançado em 1997; o periódico *Data Science Journal*, lançado em 2002; o livro *Competing on analytics*, em 2005, de Thomas H. Davenport, Don Cohen e Al Jacobson; o artigo *Rise of the data scientist*, em 2009, de Nathan Yau; o artigo *What is Data Science?*, em 2010, de Mike Loukides; o artigo *Data scientist: the sexiest job of the 21st century*, em 2012, de Thomas H. Davenport e D. J. Patil; entre outros (PRESS, 2013). Assim sendo, a busca progressiva da ligação entre métodos estatísticos tradicionais com as tecnologias

desse termo já no ano de 1996, quando os membros da "International Federation of Classification Societies (IFCS)" decidem utilizá-lo no título da conferência daquele ano: *Data science, classification, and related methods*. Desta maneira, afirmar a data exata, assim como quem proferiu o termo pela primeira vez, mostra-se não ser consenso absoluto. No entanto, existe a concordância de que o reconhecimento dessa nova área ocorre em meados da década de 1990 e início dos anos 2000, motivado, fundamentalmente, pela percepção da necessidade de compreender o recente fenômeno de grande produção de dados e, consequentemente, apresentar soluções de tratamento e uso para estes.

computacionais, que estão em constante evolução, mostra ser algo bastante característico dessa genealogia da ciência de dados.

Em relação aos principais desafios de pesquisa em ciência de dados, segundo Maneth e Poulovassilis (2017), eles incluem: o desenvolvimento de técnicas computacionais capazes de escalonar os volumes e variedades de dados que são gerados por meio de tecnologias baseadas em *Web*, móveis e difusas; a proporção de dados que estão sendo produzidos por empresas de grande porte; as aplicações científicas e de mídias sociais; o desenvolvimento de ferramentas de limpeza, transformação, modelagem, análise, integração e visualização de dados, permitindo aos cientistas de dados entenderem e melhorarem a veracidade do *big data* e extrair valor com maior rapidez, facilidade e fiabilidade; e, por fim, garantir a segurança, privacidade e propriedade de dados das organizações e dos usuários.

No que diz respeito à aplicação desse fenômeno recente em áreas tradicionais da sociedade, é cada vez mais perceptível o surgimento gradual de novas iniciativas que buscam tirar maior proveito desse campo. No setor governamental, por exemplo, como asseveram Ziviani, Porto e Ogasawara (2015), existe grande profusão de bases de dados que podem possibilitar oportunidade para análise das atividades desenvolvidas pelo setor público, objetivando tornar o planejamento mais eficiente, além de criar novos serviços que melhorem o relacionamento com o cidadão.

Por sua vez, existem empreendimentos na iniciativa privada que buscam trazer um conjunto de vantagens de negócio através da análise de dados. Isso ocorre não apenas com os dados produzidos pelas próprias entidades, mas também com dados que são comercializados com terceiros, visando agregar maior valor aos serviços e produtos que serão desenvolvidos, como apontado por Loukides (2010, p. 2, tradução nossa), quando afirma que "a questão enfrentada por toda empresa hoje, *startups*, organizações sem fins lucrativos [...] que desejam atrair uma comunidade é como usar os dados de maneira eficaz, não apenas seus próprios dados, mas todos os dados que estão disponíveis e são relevantes". Um exemplo disso é o comércio eletrônico, *e-commerce*, onde diversas empresas compram dados de navegação de usuários da internet com o objetivo de desenvolver *marketing* direcionado, baseado no histórico de navegação desses usuários.

Desta maneira, dentro desse contexto, uma definição pontual da ciência de dados é a feita por Loukides (2010, p. 1, tradução nossa, grifo nosso), afirmando que "apenas o uso de dados não é exatamente o que se quer dizer com 'ciência de dados'. Uma aplicação de dados adquire seu valor dos dados em si, criando mais dados como resultado. Não é apenas uma

aplicação com dados; é um produto de dados. *A ciência de dados permite a criação de produtos de dados*". Ou seja, a finalidade máxima da ciência de dados é permitir que sejam desenvolvidas aplicações que utilizem dados como insumos na geração de produtos e serviços para as organizações. Assim sendo, resta a indagação de quais seriam as características e requisitos que representam um profissional apto a trabalhar na construção desses produtos e serviços.

6.1 O Cientista de Dados

Esse ambiente, o *big data*, vivenciado com maior clareza nas últimas duas décadas, assim como o entendimento tanto corporativo – o viés empresarial –, quanto acadêmico – o viés do conhecimento – sobre a necessidade de uma área que fosse responsável por trazer soluções para os problemas oriundos da enorme produção de dados, propiciou o surgimento de um novo tipo de profissional responsável por desenvolver produtos e serviços a partir desses dados, designado de *cientista de dados*. Segundo Miller (2013), esses cientistas são os mágicos da era do *big data*. Eles analisam os dados utilizando modelos matemáticos e criam narrativas ou visualizações que consigam explicá-los, e depois sugerem como usar as informações para tomar decisões.

Por sua vez, Davenport e Patil (2012) discorrem que o cientista de dados é um profissional de alto nível, com treinamento e curiosidade suficientes para conseguir efetuar descobertas no mundo do *big data*. Segundo os autores, o termo foi cunhado em 2008 por D. J. Patil e Jeff Hammerbacher, respectivos líderes das iniciativas sobre análise de dados no LinkedIn e Facebook.

Ainda de acordo com Davenport e Patil (2012), o que os cientistas de dados fazem é realizar descobertas enquanto "nadam em dados". Estando à vontade no mundo digital, eles conseguem estruturar grandes quantidades de dados que não possuem forma, tornando possível analisá-los. Esses profissionais são capazes de encontrar ricas fontes de dados, conectando-as a outras fontes que estão incompletas (DAVENPORT; PATIL, 2012), propiciando, assim, que trabalhem em conjunto para alcançar os objetivos previamente traçados.

Para Davenport (2014, p. 85),

> A função do cientista de dados começou a se difundir nas organizações no final dos anos 2000, principalmente na região da Baía de São Francisco, onde as empresas se voltaram para a exploração de dados provenientes da internet e das mídias sociais [...]. As empresas acumulavam dados e mais dados, e começou-se a desenvolver uma

> enorme variedade de novas tecnologias (muitas delas inventadas por cientistas de dados) para armazenar, processar e analisar esses dados.

Assim sendo, o cientista de dados emerge como sendo um profissional que tem como principal missão trazer clareza sobre o cenário de grande produção e acúmulo de dados. Não obstante, espera-se desse especialista um conjunto de saberes e habilidades que são essenciais no desempenho do papel para o qual é destinado.

6.1.1 Habilidades

O mercado de trabalho para o cientista de dados, assim como para um profissional de qualquer área, deseja encontrar um especialista que esteja munido de um conjunto de habilidades tidas como fundamentais. Inicialmente vislumbrava-se que todas essas competências almejadas estivessem em um único indivíduo, como discorre Davenport (2014). Todavia, a partir de determinado momento, tendo em conta a dificuldade para encontrar um profissional considerado completo e que estivesse disponível no mercado, passa a ser admitido um modelo mais realista quanto ao saber exigido desses indivíduos. Isto permite que diversos especialistas, das mais variadas áreas, unam seus conhecimentos no momento de apresentar soluções para analisar, tratar e interpretar a grande quantidade de dados existentes.

Nessa mesma linha, Granville (2013 apud DAVENPORT, 2014, p. 95) distingue os cientistas de dados em dois tipos: os *verticais* e *horizontais*. Os verticais são aqueles especialistas que possuem um profundo conhecimento em algum campo específico (cientistas da computação, estatísticos, engenheiros de *software* etc.); cada qual, sendo um *expert* de sua área, pode agregar valor a processos específicos da análise, tratamento e interpretação dos dados. Por sua vez, os horizontais são os cientistas de dados que possuem um pouco de conhecimento em cada uma das áreas que podem contribuir com a dinâmica na qual os dados estão inseridos. Devido ao abrangente entendimento que possuem, esses últimos são os profissionais mais almejados pelo mercado, conseguindo, muitas vezes, lidar individualmente com todo o processo de análise, interpretação e obtenção de resultados a partir dos dados. No entanto, como observado anteriormente, esses profissionais encontram-se escassos, em grande parte devido à falta de cursos de formação específicos para cientistas de dados nas instituições de ensino. No Brasil, por exemplo, até o ano de 2016 existiam poucas iniciativas de cursos na modalidade *lato sensu*, como asseveram Curty e Serafim (2016), o que gera prejuízo no desenvolvimento das habilidades buscadas. Muitas vezes, as próprias empresas contratantes

oferecem cursos que possibilitem maior aperfeiçoamento de seus cientistas de dados (DAVENPORT; PATIL, 2012).

Entre as habilidades buscadas em um cientista de dados, nas palavras de Davenport (2014, p. 85), estão: "[ser] um *hacker*, um cientista, um analista quantitativo, um conselheiro de confiança e um *expert* em negócios". O autor apresenta um quadro, exposto logo abaixo, em que melhor especifica cada uma dessas habilidades:

Quadro 2 – Habilidades do cientista de dados

Hacker
• Capacidade de codificar; • Domínio de arquiteturas tecnológicas de *big data*.
Cientista
• Tomada de decisões baseada em evidências; • Improvisação; • Impaciência e inclinação à ação.
Conselheiro de confiança
• Grandes habilidades de comunicação e relacionamento; • Capacidade de elaborar decisões e entender os processos decisórios.
Analista quantitativo
• Análise estatística; • Visual *analytics*; • Aprendizado de máquina; • Análise de dados não estruturados, como texto, vídeo ou imagens.
***Expert* em negócios**
• Compreensão de como o negócio funciona e lucra; • Boa noção de onde aplicar o *analytics* e o *big data.*

Fonte: Davenport (2014, p. 86).

A partir do Quadro 2 é possível depreender que um cientista de dados ideal, de acordo com Davenport (2014), deve possuir um conjunto de características e competências que abarcam desde o entendimento razoável de programação e arquiteturas desenvolvidas especificamente para o ambiente *big data*; perpassando os princípios básicos da estatística, de extrema importância na ocasião da mineração e tratamento dos dados; chegando aos fundamentos acerca da gestão de negócios, liderança e proatividade, ou seja, conceitos advindos da administração. Isso corrobora com o entendimento da ciência de dados como um campo extremamente interdisciplinar.

Kim e Lee (2016, p. 166) traçam sua própria representação das habilidades e conhecimentos imprescindíveis para a atuação de um cientista de dados. Os autores chegaram a esse resultado realizando um levantamento de ofertas de emprego na área de ciência de dados

em três *sites* norte-americanos (Indeed.com, Monster.com e CareerBuilder.com) e identificando os requisitos exigidos pelos empregadores. A partir desse levantamento, Kim e Lee (2016) chegaram a um conjunto de habilidades que abarcam três classes: sistemas, negócios e técnicas. Por sua vez, essas três classes se subdividem em subclasses, que apontam para as competências demandadas. É possível perceber, *grosso modo*, que as habilidades assinaladas por Kim e Lee (2016) possuem uma "espinha dorsal" bastante similar àquelas identificadas por Davenport (2014), mostrando tendência às áreas de computação e estatística, além de alguns princípios da administração.

Quadro 3 – Habilidades e conhecimento de um cientista de dados

<table>
<tr><th>Sistemas</th><th>Negócios</th><th>Técnicas</th></tr>
<tr><td>Desenvolvimento</td><td>Social</td><td>Software</td></tr>
<tr><td rowspan="3">Análise;
Implementação/Teste;
Gestão de dados;
Conhecimento de diferentes tecnologias;
Desenvolvimento de metodologias;
Programação;
Operação/Manutenção;
Integração;
Documentação.</td><td>Habilidades interpessoais;
Comunicação;
Automotivação.</td><td rowspan="3">Linguagem de programação;
Banco de dados/Data Warehouse;
Plataformas Open Source;
Domínio de diferentes pacotes de software;
Visualização de dados.</td></tr>
<tr><td>Negócios</td></tr>
<tr><td>Conhecimento específico do setor/negócio;
Habilidade de análise macro;
Negócios online/e-commerce.</td></tr>
<tr><td>Solução de problemas</td><td>Gerencial</td><td>Arquitetura de redes</td></tr>
<tr><td rowspan="2">Modelagem de dados;
Análise quantitativa/estatística;
Pensamento analítico/lógico;
Criatividade/Inovação;
Capacidade para solução de problemas;
Adaptabilidade/Flexibilidade;
Capacidade estratégica.</td><td rowspan="2">Administração geral;
Organização/Liderança;
Capacidade de monitoramento e controle;
Planejamento;
Treinamento;
Gestão de mudança
Gerenciamento de projetos.</td><td>Internet;
Dispositivos de rede;
Computação em nuvem;
Arquitetura e segurança de rede.</td></tr>
<tr><td>Hardware</td></tr>
<tr><td></td><td></td><td>Dispositivos de armazenamento;
Impressoras;
Desktop/PC;
Servidores/Estações de trabalho.</td></tr>
</table>

Fonte: adaptado de Kim e Lee (2016, p. 166).

Por seu turno, Rodrigues, Nóbrega e Dias (2017) também esboçam um cenário em que são apresentadas algumas competências exigidas de um cientista de dados. Percebe-se que, ao contrário de Davenport (2014) e Kim e Lee (2016), os autores deram maior ênfase às

habilidades relacionadas com o campo da ciência da computação e da estatística, tidas, para muitos, como aquelas imprescindíveis para todo profissional que decida trabalhar com grandes volumes de dados.

Quadro 4 – Competências esperadas do cientista de dados

- **Capacidade de estruturar grandes volumes de dados amorfos;**
- **Tornar os dados possíveis para análise;**
- **Identificar fontes de grandes volumes de dados e cruzar com outras fontes;**
- **Criar ferramentas e analisar grande quantidade de dados;**
- **Domínio de ferramentas que deem conta do volume de dados (Hadoop, por exemplo);**
- **Formação em qualquer área, desde que tenha foco em dados e na computação.**

Fonte: adaptado de Rodrigues, Nóbrega e Dias (2017, p. [10-11]).

O último ponto, anotado por Rodrigues, Nóbrega e Dias (2017), diz respeito a formação que se espera de um profissional que decida trabalhar com dados. Essa matéria sustentada pelos autores, assim como por Davenport (2014, p. 100), indica que muitos profissionais que atuam na área de dados possuem formação não necessariamente em ciência de dados, mas em campos correlatos, como a ciência da computação e estatística, até por conta da ainda escassa oferta de cursos que sejam específicos, como já exposto anteriormente. Diante desse cenário, é relevante buscar compreender como se dá a formação dos profissionais aptos a exercer as funções de um cientista de dados.

6.1.2 Formação

No seu livro *Big data no trabalho*, Davenport (2014, p. 99) discorre que, apesar de muitos cientistas de dados possuírem formação acadêmica e títulos, "a academia nem sempre é o espaço mais eficiente para o desenvolvimento das habilidades necessárias [para atuação como cientista de dados]". O autor, com suas palavras, não busca questionar a capacidade do meio acadêmico na formação de profissionais qualificados, muito pelo contrário, apenas salienta a insípida trajetória educacional que ainda existe em relação à ciência de dados.

Quanto a isto, Curty e Serafim (2016), buscando maior compreensão sobre a matéria, efetuaram um estudo acerca da formação profissional de cientistas de dados nos Estados Unidos, conseguindo identificar um total de 93 programas, ofertados por instituições de ensino

superior, que se relacionam com a temática "ciência de dados". Isso revela maior compreensão das instituições formadoras desse país acerca da importância crescente em relação à área de análise de dados e seus benefícios para as organizações.

Os autores ainda realizaram uma análise sobre quais as exigências mínimas que as instituições cobram daqueles interessados nos cursos ofertados, tendo como base os 93 programas. O resultado alcançado foi de que 43% dos programas exigiam algum tipo de formação em nível de bacharelado e/ou cursos e créditos efetuados na área de estatística, 37,5% em matemática e 36% em ciência da computação (CURTY; SERAFIM, 2016, p. 322). Além disso, foi possível identificar também que, em muitos casos, exigia-se habilidades relacionadas com programação, cálculo, dados quantitativos e análise estatística. Desta maneira, percebe-se, novamente, a extrema relação que existe entre a ciência de dados e as áreas de computação, matemática e estatística, posto que, estas últimas, contribuíram de forma elementar, por meio de fundamentação teórica, para os princípios que deram origem à primeira, como enfatizam Curty e Serafim (2016).

Outro aspecto explorado no estudo de Curty e Serafim (2016, p. 324) foram as competências que as instituições de ensino buscam desenvolver nos egressos dos seus programas em ciência de dados. Os autores detectaram alguns eixos temáticos com forte ênfase, tais como: análise estatística avançada, aprendizagem de máquina, computação aplicada e programação, identificação de padrões e *insights* por mineração de dados, modelagem de dados, visualização e representação gráfica de dados, entre outros.

Por sua vez, colocando o Brasil em posição de análise, as iniciativas de formação em ciência de dados ainda são bastante insipientes, como mencionado anteriormente. Algumas instituições já ofertam cursos *lato sensu*, como a Universidade Presbiteriana Mackenzie, a Escola Superior de Propaganda e Marketing e a Fundação Getúlio Vargas (BRETERNITZ; LOPES; SILVA, 2015 apud CURTY; SERAFIM, 2016). No entanto, apenas recentemente surgiu o primeiro curso de graduação (tecnólogo), oferecido pelo Centro Universitário Metodista Izabela Hendrix, segundo Dino (2018).

Analisando a matriz curricular do curso ofertado pela instituição Izabela Hendriz[6], percebe-se total alinhamento com a estrutura predominante dos programas estudados por Curty e Serafim (2016), tal e qual com as habilidades expostas por Davenport (2014), Kim e Lee (2016) e Rodrigues, Nóbrega e Dias (2017), buscando desenvolvimento de competências

[6] Informações disponíveis em: <http://izabelahendrix.edu.br/ciencias-de-dados/matriz-curricular/ingressantes-no-1o-semestre>. Acesso em: 6 abr. 2018.

relacionadas com áreas da ciência da computação, estatística, matemática e administração. Algumas disciplinas que merecem destaque são: programação funcional e orientada a objetos, probabilidade e estatística, banco de dados relacionais, distribuídos e não convencionais, algoritmo e estrutura de dados, *analytics*, gestão de projetos, *machine learning*, *data mining*, *big data* e *business intelligence*.

Partindo-se do que já foi exposto até o momento, surge a indagação acerca do papel da CI como disciplina contribuinte para a ciência de dados, assim como a forma com que o profissional da informação – mais especificamente, o bibliotecário – pode colaborar e, mesmo, atuar como cientista de dados. Viu-se, até aqui, que a ciência de dados está fortemente inserida dentro de disciplinas do ramo das tecnologias e das exatas (ciência da computação, estatística e matemática) e, em certo nível, da administração. Desta maneira, num primeiro momento, associar essa área com a CI soa um tanto ilógico, posto que, enquanto uma engloba princípios das ciências exatas/tecnologias, a outra está unida às ciências sociais. Todavia, tal afirmação mostra-se equivocada. Apesar das diferenças inquestionáveis entre as duas disciplinas, existe, sim, espaço em que as semelhanças se apresentam. Como demonstrado anteriormente, a razão de ser da CI fundou-se na recuperação da informação, uma vez que, as fontes informacionais tiveram um salto significativo no pós-guerra de 1945. Com a ciência de dados não ocorreu de forma distinta, posto que, a partir do maior desenvolvimento tecnológico ocorrido nas últimas décadas, viu-se crescer em conjunto o volume de dados e, com isto, a necessidade de uma área responsável por entender e encontrar soluções para tal fenômeno.

Deste modo, sendo a informação o produto dos dados, visto que a informação é o dado colocado dentro de um contexto (ANGELONI, 2003), é de grande relevância para o profissional da informação a busca pelo estreitamento dos vínculos com a ciência de dados. A esse especialista cabe rastrear os requisitos necessários para poder atuar como cientista de dados, até porque, como defendido por Davenport (2014) e Rodrigues, Nóbrega e Dias (2017), não é necessário um diploma em área específica para ser reconhecido como um cientista de dados, mas, sim, determinadas competências tidas como essenciais.

7 METODOLOGIA DE PESQUISA

A metodologia utilizada neste estudo possui abordagem qualitativa, uma vez que "considera [haver] uma relação dinâmica entre o mundo real e o sujeito, isto é, um vínculo indissociável entre o mundo objetivo e a subjetividade do sujeito que não pode ser traduzido em números" (PRODANOV; FREITAS, 2013, p. 70). Constitui-se, também, como uma pesquisa exploratória, posto que, objetiva estabelecer a existência de vínculo entre os conceitos analisados, neste caso, o fenômeno *big data*, o novo campo que emerge, a ciência de dados, e o profissional da informação. Além disso, investiga as habilidades e competências que o profissional da informação possui em semelhança com o cientista de dados, e se a formação acadêmica do primeiro o qualifica para atuar na área de ciência de dados. Gil (2002, p. 41) salienta que o objetivo principal da pesquisa exploratória é "proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses". O autor também destaca como sendo as técnicas de coleta de dados mais utilizadas nesse tipo de metodologia a pesquisa bibliográfica e o estudo de caso.

Quanto ao procedimento adotado, este trabalho utiliza como método a pesquisa bibliográfica. O levantamento dos dados ocorreu a partir de uma pesquisa bibliográfica sobre os termos "ciência da informação", "information science", "profissional da informação", "information professional", "big data", "ciência de dados", "data science", "cientista de dados" e "data scientist", realizada junto ao Google Acadêmico e as bases de dados Scielo, Lisa e Brapci, entre os dias 19/12/2017 e 26/04/2018. Para Marconi e Lakatos (2003, p. 183) uma pesquisa bibliográfica, também conhecida como de fontes secundárias, tende a analisar os materiais que já foram produzidos sobre determinado tema de estudo, abrangendo diversos tipos de publicações, tais como livros, revistas, monografias, jornais etc. Ainda sobre a pesquisa bibliográfica, Prodanov e Freitas (2013, p. 54) destacam a importância da verificação dos dados que são obtidos, atentando-se a possíveis incoerências e contradições apresentadas pelas obras analisadas. Neste trabalho foram considerados como fontes bibliográficas artigos de periódicos e livros. Por sua vez, ao longo da pesquisa bibliográfica, a seleção das fontes utilizadas neste trabalho foi feita a partir da técnica de leitura dinâmica, do título e do resumo, como destaca Santa Anna (2017, p. 24), sendo que, os materiais não relacionados com o tema foram descartados.

Posteriormente à pesquisa bibliográfica, como mecanismo para melhor compreensão da temática, foi efetuada análise das disciplinas ofertadas dentro do curso de biblioteconomia da

Universidade de Brasília, afim de identificar aquelas que oferecem conhecimentos e desenvolvem habilidades úteis para o futuro profissional da informação no campo da ciência de dados. Essa análise foi realizada a partir do portal eletrônico da UnB chamado "Matrícula Web"[7], que, entre outras coisas, disponibiliza a oferta de disciplinas dos cursos. Após acesso aos tópicos "Curso", "Darcy Ribeiro", "Biblioteconomia" e "Currículo", passo a passo para se chegar às disciplinas do curso, foi feito, em um primeiro momento, uma análise dos títulos das matérias elencadas no portal. Depois da identificação das disciplinas consideradas pertinentes, passou-se ao estudo de suas ementas e programas. Essas etapas foram efetuadas entre os dias 29/04/2018 e 30/04/2018.

A partir do referencial teórico identificado através da pesquisa bibliográfica, assim como da análise de currículo, realizada junto ao curso de biblioteconomia da UnB, verificou-se um conjunto de habilidades que caracterizam o cientista de dados e o profissional da informação, tal qual, disciplinas relevantes na formação do especialista em informação para atuação na área de ciência de dados. Os resultados dessa análise são apresentados a seguir.

[7] Disponível em: <https://matriculaweb.unb.br/graduacao/default.aspx>. Acesso em: 3 maio 2018.

8 O PROFISSIONAL DA INFORMAÇÃO E A CIÊNCIA DE DADOS

Anteriormente, neste trabalho, foi realizada uma breve descrição acerca das habilidades e conhecimentos esperados dos profissionais que lidam constantemente com o insumo informação. Verificou-se que vários autores passam a utilizar a expressão *Moderno Profissional da Informação* (SANTOS, 1996; GUIMARÃES, 1997; VALENTIM, 2000; SANTA ANNA; PEREIRA; CAMPOS, 2014; SANTA ANNA, 2017) para se referir ao especialista que emerge em um ambiente altamente dinâmico, pautado pelo constante avanço das tecnologias de informação e comunicação, bem como pelas demandas sociais. Neste momento, surge a indagação sobre qual papel esse profissional desempenhará nesse novo cenário de grande produção de dados, em alta velocidade e variedade, assim como na *ciência de dados*, campo que aflora em conjunto.

Em seu trabalho *Big data: os novos desafios para o profissional da informação*, Ribeiro (2014) defende que o profissional da informação busque seu lugar dentro dessa nova dinâmica informacional surgida com o advento do *big data*. Para o autor, esse especialista deve estar atento às pesquisas relacionadas com os grandes volumes de dados, em especial "nos temas iniciais desenvolvidos na fase de *Discovery*, *Data Preparation* e *Modeling Planning*" (RIBEIRO, 2014, p. 102). Em outras palavras, o profissional da informação precisa estar mais preocupado em ocupar-se com processos do que com técnicas, posto que suas atribuições estão fundadas na forma que se dá o fluxo e o consumo da informação, como discorre Ferreira (2003).

Indo além, Coneglian, Gonçalvez e Santarém Segundo (2017) argumentam que o profissional da informação pode assumir um papel maior na análise de dados, mais especificamente, em processos técnicos do tratamento dos dados. Baseando-se em uma proposta de Bugembe (2016 apud CONEGLIAN; GONÇALVEZ; SANTARÉM SEGUNDO, 2017) acerca da análise de dados, são identificadas as etapas da análise de dados nas quais o profissional da informação pode trazer contribuições. A seguir são apresentadas essas etapas assinaladas pelos autores, assim como suas respectivas atividades.

Quadro 5 – Etapas da análise de dados e o papel do profissional da informação

Etapas	Papel do profissional da informação
• Fontes dos dados	• Determinar as fontes utilizadas (internamente e externamente).
• Captura e armazenamento dos dados	• Avaliar (o que capturar e quando capturar).

• Processamento e fusão dos dados	• Estruturação dos dados (representação e formas de estruturar).
• Acesso aos dados	• Disponibilizar os dados (promover acesso a outras áreas).
• Análise dos dados	• Participação (identificar as necessidades dos gestores).
• Exposição dos dados	• Apresentar os resultados (interface com o gestor).

Fonte: elaborado pelo autor a partir de Coneglian, Gonçalvez e Santarém Segundo (2017, p. 137).

Para os autores, o profissional da informação possui longo conhecimento na identificação das melhores fontes existentes, tendo a capacidade de trabalhar com fontes de qualquer natureza. Seguem afirmando que, na fase de captura e armazenamento dos dados, trabalhando em colaboração com cientistas da computação, o profissional da informação delineia quais são os melhores dados para a análise, sendo favorecido nessa etapa por ter maior conhecimento das fontes utilizadas. Os autores defendem, em relação a etapa de processamento e fusão, que o profissional da informação poderá atuar no processo de estruturação dos dados, posto que possui conhecimento suficiente para auxiliar na conversão de massas de dados em informações estruturadas; esse conhecimento pertinente vai desde estruturas chave-valor, metadados, passando pela representação da informação. Por sua vez, na fase de acesso aos dados, Coneglian, Gonçalvez e Santarém Segundo (2017) afirmam que o profissional da informação servirá como interlocutor entre o gestor da organização e o processo que está sendo realizado, buscando definir quais dados serão tratados e se há algum tipo de mudança. Quanto à etapa de análise dos dados, os autores defendem que esta é atribuição exclusiva do cientista da computação, responsável em construir os algoritmos que serão utilizados na análise e pela definição das técnicas estatísticas usadas, cabendo ao profissional da informação atuar de forma secundária no direcionamento dos rumos da análise, assim como das variáveis que serão consideradas. Por último, na fase de exposição dos resultados, é defendido que tanto o profissional da informação quanto o cientista da computação apresentem os resultados obtidos com a análise, competindo ao primeiro o papel principal, "por conseguir expor com mais propriedade os resultados atingidos" (CONEGLIAN; GONÇALVEZ; SANTARÉM SEGUNDO, 2017, p. 138).

A partir da análise feita a respeito das habilidades almejadas em um profissional da informação e aquelas esperadas de um cientista de dados é possível traçar um quadro comparativo entre esses dois profissionais.

Quadro 6 – Habilidades de um cientista de dados e de um profissional da informação identificadas na literatura

Habilidades do cientista de dados	Habilidades do profissional da informação
Computação	**Técnico-científicas**
• Linguagens de programação (Python, JavaScript etc.); • Tecnologias em *big data* (Hadoop etc.); • Banco de dados relacionais e não relacionais (NoSQL etc.); • *Machine learning*; • *Data mining*; • *Analytics*; • Computação em nuvem; • Visualização de dados; • Segurança de dados.	• Domínio na utilização de equipamentos eletrônicos e de softwares específicos; • Conhecimento de bases de dados; • Tratamento de fontes de informação em diferentes suportes, unidades e serviços de informação.
Quantitativas/Estatística	**Gerenciais**
• Análise quantitativa/estatística; • *Data mining*; • *Analytics*; • Visualização de dados.	• Conhecimento do ambiente de negócios; • Capacidade de trabalhar em grupo; • Familiaridade na administração de *info-business*; • Capacidade de gerenciar episódios de diferentes naturezas; • Habilidade na identificação de parceiros; • Compreensão das necessidades informacionais do administrador; • Identificação de informações relevantes.
Gerenciais	
• Organização/Liderança; • Capacidade para solução de problemas; • Capacidade estratégica; • Compreensão do negócio; • Tomada de decisões.	
Sociais	**Comunicação, Sociais e Políticas**
• Habilidades interpessoais; • Boa comunicação; • Automotivação.	• Excelência na comunicação oral e escrita; • Atualização profissional constante; • Desenvolvimento de produtos voltados para a comunicação entre usuários e as fontes informacionais.

Fonte: elaborado pelo autor a partir de Davenport (2014), Kim e Lee (2016), Valentim (2000) e Ferreira (2003).

Partindo-se do que é apresentado no Quadro 6 é possível identificar que as habilidades esperadas de um cientista de dados e aquelas almejadas num profissional da informação possuem similaridades, em certos aspectos. Isso pode ser verificado, principalmente, nas competências de caráter Gerenciais, Sociais e Políticas e de Comunicação. No entanto, também

é possível constatar que existem diferenças significativas entre os dois profissionais. Nota-se, por exemplo, maior exigência de conhecimento sobre Computação e Estatística para o cientista de dados do que para o profissional da informação. Esse fato corrobora com a ideia de Coneglian, Gonçalvez e Santarém Segundo (2017), quando defendem que a atividade de análise dos dados, no ambiente *big data*, é de atribuição do cientista da computação, responsável pelo desenvolvimento de algoritmos e a implantação de princípios estatísticos, ficando a cargo do profissional da informação a tarefa de auxiliá-lo quanto a isto. Além disso, o fato também confirma a percepção de alguns autores (DAVENPORT, 2014; CONEGLIAN; GONÇALVEZ; SANTARÉM SEGUNDO, 2017) de que a profissão de cientista de dados não diz respeito a um único indivíduo munido com todo o conhecimento necessário para realizar a análise e interpretação dos dados, mas sim, de diversos profissionais contribuindo mutuamente.

Não obstante, ainda na busca de maior entendimento de como se dá a relação entre a ciência de dados e o profissional da informação, é pertinente realizar uma análise de currículo de curso que se encaixa dentro do escopo da CI. Isso permitirá identificar as disciplinas ofertadas que possuem algum tipo de correspondência com a ciência de dados. Por motivos de proximidade com o autor deste trabalho, o curso escolhido para análise é o de biblioteconomia da UnB.

8.1 O Curso de Biblioteconomia da UnB

Como mencionado anteriormente nesta pesquisa, a abordagem de trabalho sobre o profissional da informação é direcionada para o profissional bibliotecário. Por esse motivo, a análise de currículo de curso associado com a CI, aqui proposto, é direcionada para o curso de biblioteconomia, mais especificamente o da UnB.

O curso de biblioteconomia da UnB é uma graduação em nível de bacharelado, previsto para ser concluído ao longo de 8 semestres (4 anos). Sendo o profissional apto a trabalhar com informação, não importando o espaço na qual está inserida. Além disso, o profissional bibliotecário pode atuar em diferentes locais, como centros de informação, institutos de pesquisas, bibliotecas, provedores de informação, empresas, bancos, centros culturais etc.[8], ou seja, em todo ambiente que tenha a informação como insumo. Em relação às diretrizes do curso, é possível destacar duas:

[8] Informações disponíveis em: <http://biblioteconomia.fci.unb.br/index.php>. Acesso em: 29 abr. 2018.

> [...] oferecer um meio ambiente de imersão integral nas tecnologias da informação, enquanto instrumentos convencionais da ação profissional do bibliotecário;
> caracterizar a informação como todo conhecimento humano inscrito sob qualquer forma, e considere a gestão desses recursos do universo da informação far-se-á independentemente de sua origem, de seu suporte material e de sua instituição depositária[9].

No tocante à quantidade de créditos exigidos, o estudante da graduação em biblioteconomia da UnB necessita alcançar 180 créditos para obter o diploma de bacharel. Esses créditos estão divididos em 7 grupos de disciplinas, a saber: *disciplinas obrigatórias* (22 disciplinas, totalizando 100 créditos); *disciplinas da cadeia 1* (8 opções de disciplinas, sendo que o estudante precisa cursar apenas uma ou fazer 4 créditos); *disciplinas da cadeia 6* (3 opções de disciplinas, sendo que o estudante precisa cursar apenas uma ou fazer 4 créditos); *disciplinas da cadeia 7* (4 opções de disciplinas, sendo que o estudante precisa cursar apenas uma ou fazer 4 créditos); *disciplinas da cadeia 8* (2 opções de disciplinas, sendo que o estudante precisa cursar apenas uma ou fazer 4 créditos); *disciplinas da cadeia 9* (2 opções de disciplinas, sendo que o estudante precisa cursar apenas uma ou fazer 4 créditos); e *disciplinas optativas* (177 opções de disciplinas para o estudante completar a quantidade de créditos necessária para concluir o curso)[10]. Além desses 7 grupos, também existe a opção de *disciplinas de módulo livre*, sendo disciplinas que não estão no fluxo do curso, nem são optativas, mas que o estudante pode cursar, ou seja, qualquer disciplina ofertada pela universidade, desde que o estudante preencha os pré-requisitos exigidos. Nessa última opção, o discente pode cursar o limite máximo de 24 créditos.

A partir do reconhecimento da matriz curricular do curso de biblioteconomia da UnB, buscou-se identificar quais disciplinas apresentam alguma similaridade com as atribuições e competências esperadas de um cientista de dados, verificadas no referencial teórico deste trabalho. Para tanto, foi realizada uma análise a respeito das ementas das disciplinas, assim como daquilo que elas propõem em seus respectivos programas pedagógicos.

[9] Idem.
[10] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/curriculo.aspx?cod=8222>. Acesso em: 30 abr. 2018.

8.1.1 Estatística Aplicada

A disciplina de Estatística Aplicada é uma matéria de 6 créditos obrigatória para o curso de biblioteconomia da UnB. Essa matéria não possui pré-requisito para ser cursada, sendo ofertada no primeiro semestre da graduação. Na ementa da disciplina é indicado o desenvolvimento de conceitos básicos sobre "distribuição de frequências e suas características", "introdução à probabilidade – ajustamento de funções reais – correlação e regressão linear" e "noções de amostragem e testes de hipótese"[11]. Por sua vez, quanto ao programa da disciplina, há apontamentos para o aprofundamento de assuntos como a estatística indutiva e dedutiva, análise gráfica, média aritmética, moda e mediana, variância, desvio padrão e coeficiente de variação, variável aleatória, distribuição de probabilidade, coeficiente de correlação linear, amostras aleatórias, distribuição de amostragem etc., ou seja, conhecimentos imprescindíveis para a compreensão da estatística.

O conhecimento em estatística, como indicado por Davenport (2014) e Kim e Lee (2016), é uma das mais importantes habilidades esperadas num cientista de dados, uma vez que, a estatística, como apontado anteriormente, foi um dos pilares para o surgimento da ciência de dados. Ser apresentado aos conceitos básicos dessa disciplina possibilita ao profissional da informação a compreensão, mesmo que de forma abstrata, da dinâmica existente na atividade de análise quantitativa, bem como dos princípios que a circundam, por exemplo, o *analytics* e a mineração de dados.

8.1.2 Informática Documentária

Esta disciplina também é obrigatória para o egresso do curso, sendo de 4 créditos, é geralmente oferecida no sétimo semestre do curso. Para o discente ser matriculado nessa disciplina ele precisa ter cursado a matéria de Planejamento e Elaboração de Bases de Dados ou Arquivo Corrente 1. Em relação à ementa da disciplina, é proposto para o estudante o "uso das tecnologias e métodos relacionados com a informática aplicada aos processos documentários", "princípios de análise funcional", "automação de serviços de informação" e "automação dos processos de bibliotecas"[12]. Em seu turno, o conteúdo programático da matéria

[11] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=115011>. Acesso em: 30 abr. 2018.
[12] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=182508>. Acesso em: 30 abr. 2018.

está fundado em conceitos sobre usuário, usabilidade de sistemas de informação e acessibilidade, linguagens de marcação, ou seja, metalinguagens (HTML, XML) e arquivos eletrônicos. Todavia, também existe orientações sobre conceitos básicos de linguagens de programação, como por exemplo, a linguagem Python, passível de utilização em ambientes *big data*, conforme Davenport (2014).

É importante destacar que, assim como o conhecimento sobre estatística, os princípios sobre computação e as novas tecnologias baseadas em *big data* não são de atribuição do profissional da informação. Essas competências são buscadas em um cientista da computação e/ou estatístico. Cabe ao especialista em informação ter conhecimento razoável sobre essas áreas, visando auxiliar no processo de análise dos dados, de forma complementar, como defendem Coneglian, Gonçalvez e Santarém Segundo (2017).

8.1.3 Planejamento e Elaboração de Bases de Dados

A disciplina de Planejamento e Elaboração de Bases de Dados também é obrigatória para o curso de biblioteconomia da UnB. A matéria oferece 4 créditos ao graduando, sendo ofertada normalmente no quarto semestre. Possui três opções de pré-requisitos, sendo: ter cursado as disciplinas de Catalogação e Introdução à Microinformática ou Catalogação e Introdução à Ciência da Computação ou Catalogação e Introdução ao Processamento de Dados. Na sua ementa há proposta de desenvolver conceitos sobre "caracterização de bases de dados", "estudo de viabilidade", "planejamento, projeto e implantação de bases de dados bibliográficas" etc[13]. Quanto ao programa, são apresentados assuntos relacionados com os tipos de bases de dados, estruturas e formatos de dados, gerenciadores de bases de dados, modelos lógicos, coleta e conversão de dados, atualização dos dados, escolha de formatos, *hardware* e *software*, entre outros.

Outro tópico de grande relevância dentro da coleta de vastos volumes de dados. O domínio sobre a matéria de bases de dados, como verificado em Ferreira (2003), está entre as maiores competências esperadas de um profissional da informação pelo mercado. De fato, o aprendizado de habilidades relacionadas com a identificação de fontes de dados confiáveis, assim como princípios de criação de bases de dados e organização dos dados, visando seu uso

[13] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=182877>. Acesso em: 30 abr. 2018.

futuro, pode se caracterizar como uma das grandes contribuições que o profissional da informação pode dar ao ambiente *big data* e, consequentemente, à ciência de dados.

8.1.4 Análise da Informação

Esta disciplina também é de caráter obrigatório para o curso, oferecendo 4 créditos ao graduando. Ela é comumente oferecida no terceiro semestre do bacharelado, possuindo como pré-requisito a disciplina de Controle Bibliográfico ou Museologia 1 ou Arquivo Corrente 2. Em sua ementa verifica-se a proposta de "conceituação, fundamentação teórica, objeto e função da análise da informação", "contextualidade da informação no ciclo documentário", "linguística, semântica, semiótica", "métodos e técnicas da análise da informação", "resumo documentário e indexação" etc[14]. Por sua vez, em se tratando do programa da disciplina, é possível identificar conteúdos relacionados com o ciclo documentário, as linguagens documentárias, descrição temática da informação, mineração de dados e de textos, busca e recuperação da informação, valor agregado da informação, entre outros.

O conteúdo ministrado nessa disciplina compartilha tópicos de grande pertinência para a ciência de dados. O assunto "busca e recuperação da informação" pode trazer contribuições relevantes para o processo de obtenção de dados significativos de fora da empresa. Por sua vez, os princípios das linguagens documentárias podem ser utilizados em modelos de arquitetura da informação, como verificado em Victorino et al. (2017), úteis nas aplicações de dados. Por fim, cabe destaque ao assunto "mineração de dados", conceito advindo da estatística muito pertinente dentro da análise de grandes volumes de dados, conforme Han, Kamber e Pei (2012), uma vez que possibilita a separação de dados considerados relevantes daqueles que não possuem capacidade de trazer vantagem competitiva para a organização.

8.1.5 Outras Disciplinas

Além das disciplinas do curso de biblioteconomia da UnB, expostas acima, identificadas com maior nível de relevância dentro da perspectiva de formação de um profissional apto a atuar como cientista de dados, existem outras matérias que também possuem certo grau de similaridade com as habilidades esperadas de um profissional que trabalha com dados.

[14] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=182401>. Acesso em: 30 abr. 2018.

A disciplina de *Introdução à Administração* constitui-se como também sendo matéria obrigatória para o curso, ofertando 4 créditos ao graduando. Não existem pré-requisitos para cursá-la, sendo oferecida normalmente no segundo semestre do curso. Na ementa da disciplina é possível identificar a busca de conceitos relacionados com o "planejamento, organização, liderança e controle", "estruturas organizacionais", "funções administrativas", "perspectivas da administração na sociedade contemporânea" etc[15]. Em se tratando do programa da disciplina, há proposta de entendimento de temas como poder e liderança, controle organizacional, processo decisório e resolução de problemas, a função da organização, entre outros. Tanto Davenport (2014) quanto Kim e Lee (2016) destacam entre as habilidades de um cientista de dados a necessidade de conhecer a estrutura organizacional da empresa, saber trabalhar em equipe, possuir liderança e competência para resolver problemas, todas características passíveis de serem trabalhadas nesta disciplina.

Outra disciplina que merece destaque é a de *Introdução à Ciência da Computação*. Essa matéria é optativa para o discente de biblioteconomia da UnB. Não possui pré-requisitos para ser cursada, ofertando 4 créditos ao graduando. Na ementa da disciplina pode ser visualizada a proposta de desenvolver conceitos a respeito do "pensamento computacional", variáveis e entrada de dados", "estrutura de dados-agregados homogêneos e heterogêneos" etc[16]. Quanto ao programa, verifica-se a abordagem de assuntos relacionados ao pensamento computacional, estruturas condicionais, estruturas de repetição, funções/procedimentos, arquivos, bibliotecas, entre outros. Como já mencionado, a área da computação é um dos cernes da ciência de dados, servindo de alicerce para seu surgimento. Portanto, para um profissional que busca espaço nesse ambiente é fundamental desenvolver competências, mesmo que introdutórias, a respeito dos paradigmas computacionais.

Por fim, destaca-se a disciplina de *Banco de Dados*, que também é optativa para o estudante de biblioteconomia. A matéria possui como pré-requisito a disciplina de Organização de Arquivos, e oferece 4 créditos ao discente. Em relação a sua ementa, a disciplina preocupa-se em desenvolver no estudante níveis de conhecimento sobre "abstração de informações e dados", "modelo relacional, de rede e hierárquico", "segurança, integridade e concorrência" etc[17]. Já em relação ao seu programa, a matéria propõe atividades referentes a linguagens de

[15] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=181013>. Acesso em: 30 abr. 2018.
[16] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=113913>. Acesso em: 30 abr. 2018.
[17] Informações disponíveis em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=116378>. Acesso em: 30 abr. 2018.

descrição e manipulação de dados, violações de segurança, restrições de integridade, bancos de dados não convencionais, bancos de dados orientados a objetos, bancos de dados distribuídos, entre outras. Esses saberes que a disciplina se propõe a explorar são de grande valia para a ciência de dados e o ambiente *big data*, em especial o conteúdo relacionado com bancos de dados não convencionais. O banco de dados NoSQL, muito utilizado para atuar junto a grandes volumes de dados (VICTORINO et al., 2017; LOUKIDES, 2010), está inserido nesse grupo de bancos de dados não convencionais.

8.2 As Competências do Cientista de Dados e as Disciplinas do Curso de Biblioteconomia da UnB

A partir da identificação de disciplinas, obrigatórias e optativas, do curso de biblioteconomia da UnB, que estão alinhadas, de alguma forma, com as habilidades que se propõe desenvolver em um cientista de dados, as quais foram verificadas na literatura de Davenport (2014) e Kim e Lee (2016), é possível visualizar um relacionamento, mesmo que simplificado, entre as competências de um cientista de dados e as disciplinas do curso da UnB.

Quadro 7 – Interseção entre as habilidades esperadas do cientista de dados e as disciplinas selecionadas do curso de biblioteconomia da UnB

Habilidades do cientista de dados	Disciplinas do Curso de Biblioteconomia da UnB
Computação	
• Linguagens de programação (Python, JavaScript etc.); • Tecnologias em *big data* (Hadoop etc.); • Banco de dados relacionais e não relacionais (NoSQL etc.); • *Machine learning*; • *Data mining*; • *Analytics*; • Computação em nuvem; • Visualização de dados; • Segurança de dados.	• *Informática Documentária* (linguagem Python, HTML, XML etc.); • *Estatística Aplicada* (análise gráfica, amostras aleatórias, distribuição de amostragem, etc.); • *Análise da Informação* (mineração de dados e de textos etc.); • *Planejamento e Elaboração de Bases de Dados* (estruturas e formatos de dados, modelos lógicos, coleta e conversão de dados etc.); • *Introdução à Ciência da Computação* (estrutura de dados-agregados homogêneos e heterogêneos, estruturas de repetição etc.);

	• *Banco de Dados* (bancos de dados convencionais e não convencionais etc.).
Quantitativas/Estatística	
• Análise quantitativa/estatística; • *Data mining*; • *Analytics*; • Visualização de dados.	• *Estatística Aplicada* (análise gráfica, amostras aleatórias, distribuição de amostragem, etc.); • *Análise da Informação* (mineração de dados e de textos etc.).
Gerenciais	
• Organização/Liderança; • Capacidade para solução de problemas; • Capacidade estratégica; • Compreensão do negócio; • Tomada de decisões.	• *Introdução à Administração* (liderança, controle organizacional, processo decisório e resolução de problemas, função da organização etc.).
Sociais	
• Habilidades interpessoais; • Boa comunicação; • Automotivação.	• *Introdução à Administração* (liderança).

Fonte: elaborado pelo autor a partir de dados da pesquisa.

Entre os quatro conjuntos de habilidades esperadas de um cientista de dados, a *computação* aparece como a que tem maior abrangência com as disciplinas selecionadas. Essa associação acontece tanto diretamente, através de disciplinas que buscam tratar, especificamente, dos conceitos e paradigmas tecnológicos (Banco de Dados e Introdução à Ciência da Computação, por exemplo), quanto indiretamente, por meio de matérias que abordam, de maneira superficial, conceitos da computação ou que emprestam teorias e conceitos a esse campo (Estatística Aplicada, por exemplo). Em seguida surge a área das análises *quantitativas/estatísticas*, possuindo uma disciplina específica que trata de assuntos relacionados com a estatística descritiva e inferencial (Estatística Aplicada), assim como uma disciplina que utiliza de matéria alusiva a conceitos estatísticos, a mineração de dados, (Análise da Informação). Em se tratando das habilidades *gerenciais*, a disciplina de Introdução à Administração mostra-se possuidora de alto grau de compatibilidade com os requisitos sobre gestão de negócios, visualizados entre as competências do cientista de dados. Por último, no que diz respeito às habilidades *sociais*, novamente aparece a disciplina de Introdução à Administração, mesmo que em nível mais baixo de compatibilidade.

Apesar das disciplinas aqui analisadas terem sido identificadas como sendo aquelas que possuem mais ligação, dentro do curso de biblioteconomia da UnB, com os conhecimentos e

habilidades almejados em um cientista de dados. Não obstante, existem espaços dentro do curso da UnB para debater questões mais específicas da análise de grandes volumes de dados.

A disciplina de *Tópicos Especiais em Biblioteconomia e Ciência da Informação* pode ser caracterizada como sendo um desses espaços de debate. Em sua ementa, a disciplina propõe estudar "temas de caráter geral ou específico complementares às demais disciplinas do curso e relacionados com a Biblioteconomia ou com a Ciência da Informação"[18]. Ou seja, a partir do momento que houver maior percepção sobre a importância, tanto para a biblioteconomia quanto para a ciência da informação, de uma formação mais voltada para a questão da análise dos grandes volumes de dados e sua potencialidade de gerar informações relevantes, o estudo dos temas complementares, pelos quais a disciplina se compromete, poderá ser direcionado para o fenômeno *big data* e, consequentemente, para a ciência de dados.

Por fim, vale ressaltar que no Brasil já existem algumas graduações, voltadas para a área de ciência da informação, preocupadas em desenvolver competências que possibilitem aos seus futuros profissionais maior familiaridade com o contexto dos dados produzidos em ambiente virtual, seus desafios e oportunidades, em especial os cursos relacionados com a gestão da informação. Essas graduações podem ser usadas como modelo pelo curso de biblioteconomia da UnB, posto que dialogam, diretamente, com conceitos e práticas essenciais para o entendimento do fenômeno *big data* e da ciência de dados. Os seguintes cursos ganham destaque:

- *Gestão da Informação da PUC/Campinas*: a graduação possui disciplinas que discorrem sobre Banco de Dados, Ontologia e Taxonomia da Informação, Ciência dos Dados, Métodos Quantitativos na Informação, Inteligência Competitiva e de Negócios, Segurança da Informação etc.;
- *Gestão da Informação da UFPR/Curitiba*: o curso conta com disciplinas que tratam sobre Introdução à Lógica, Fundamentos da Matemática, Segurança da Informação, Métodos Matemáticos para Gestão da Informação, Introdução à Estatística, Modelagem de Sistemas, Banco de Dados I e II, Mineração de Dados, Métodos Quantitativos de Análise etc.;
- *Ciência da Informação da UFSC/Florianópolis*: as disciplinas do curso abarcam conteúdos relacionados com a Lógica, Banco de Dados, Gestão da Qualidade,

[18] Informação disponível em: <https://matriculaweb.unb.br/graduacao/disciplina.aspx?cod=107689>. Acesso em: 18 maio 2018.

Estatística Aplicada, Mineração de Texto, *Linked Data*, Inteligência Competitiva, *Data Science*, Introdução a Algoritmos etc.

9 CONSIDERAÇÕES FINAIS

Neste trabalho verificou-se que a área de ciência da informação surge motivada pela constatação da necessidade de compreender os processos nos quais a informação se encontra envolvida e sua real importância para o desenvolvimento sociocultural dos indivíduos. Além disso, a explosão informacional também contribuiu com o processo de nascimento dessa disciplina, iniciada em meados da década de 1940, assim como a necessidade de recuperação da informação, questão que ganha destaque por volta da década de 1960, em grande parte pelo avanço das novas tecnologias. Desta forma, a CI surge com a missão de se debruçar sobre o fluxo percorrido pela informação, buscando seu entendimento e otimização.

Essa evolução tecnológica observada também trouxe grandes avanços para o meio social, desde *hardware* (equipamentos) até *software* (lógica responsável por possibilitar o funcionamento dos equipamentos), influenciando, de maneira ímpar, a forma como os indivíduos interagem com o mundo que os circunda, com os outros e com eles mesmos (CONEGLIAN; GONÇALVEZ; SANTARÉM SEGUNDO, 2017, p. 140).

Provocado justamente por essa evolução tecnológica, a partir do barateamento do armazenamento das informações e do advento da internet e das mídias sociais, surge o conceito de *big data*. Esse fenômeno se constitui como grandes *volumes* de dados, produzidos em rápida *velocidade* e com grande *variedade* de formatos (DAVENPORT, 2014). Partindo-se desse contexto, acaba emergindo, em conjunto, a necessidade de um profissional que consiga apresentar soluções, realizar análises e extrair valor dessa grande quantidade de dados, o que, por sua vez, colabora com o surgimento de um novo campo de estudo, a chamada *ciência de dado*s. Os cientistas de dados, de acordo com Miller (2013), podem ser considerados mágicos da era do *big data*. Eles conseguem analisar os dados através de modelos matemáticos, criando narrativas e visualizações que possibilita explicá-los.

Uma das características da profissão de cientista de dados, conforme Davenport (2014), é seu viés multifacetado, sendo que diversos especialistas podem exercer essa atividade, desde cientistas da computação, estatísticos, administradores, profissionais da informação, entre outros. Isso se deve à abrangência das etapas relacionadas com a análise dos dados, sendo: identificação das fontes, captura e armazenamento, acesso, análise e, por fim, exposição dos resultados dos dados analisados (CONEGLIAN; GONÇALVEZ; SANTARÉM SEGUNDO, 2017). Todavia, destaca-se que a profissão de cientista de dados é marcada por sua natureza mais voltada a conhecimentos relacionados com a computação – linguagens algorítmicas,

tecnologias para *big data* (Hadoop), banco de dados (NoSQL), aprendizado de máquina etc. – e com a estatística – análise quantitativa, mineração de dados, visualização de dados etc., – posto que a ciência de dados surgiu pautada nessas duas disciplinas.

Todavia, como destacado anteriormente, o mercado da ciência de dados está aberto aos conhecimentos advindos de um profissional da informação. Verificou-se, por exemplo, que no curso de biblioteconomia da Universidade de Brasília esse profissional é apresentado, desde cedo, a conceitos de profunda relevância para atuar nessa carreira. Ele desenvolve atividades relacionadas com linguagens de programação (Python), bases de dados, fontes de informação (fontes de dados), análise quantitativa, organização da informação (linguagens documentárias), assim como princípios da administração, também muito destacado pelos autores como característica de um cientista de dados. Além disso, os discentes do curso de biblioteconomia da UnB dispõem de outras disciplinas, de cunho optativo, que podem auxiliar no aprofundamento de suas competências, por exemplo, Introdução à Ciência da Computação e Banco de Dados.

Ainda em relação ao curso de biblioteconomia da UnB, apesar de haver matérias, em sua grade curricular, que de fato dialogam com conceitos, da ciência de dados, pertinentes para o futuro profissional da informação, há, sim, espaço para um maior aprofundamento quanto a isso. Verificou-se, por exemplo, que outros cursos no Brasil, também relacionados com a CI, possuem uma abordagem mais voltada para o fenômeno *big data* e para a ciência de dados. Esses cursos podem servir de parâmetro para possíveis mudanças realizadas junto ao curso da UnB.

Com a pesquisa, percebe-se, também, que o profissional da informação pode contribuir com competências significativas à ciência de dados, executando tarefas em diversas etapas da análise dos dados, em especial, as de descoberta, preparação e modelagem dos dados, como apontado por Ribeiro (2014). Inclusive, a criação de modelos é algo fundamental quando o assunto são grandes quantidades de dados. Frické (2015) destaca justamente isso, que a ciência necessita se preocupar mais com problemas, pensamentos, experimentos projetados, ou seja, com mais teoria (modelo) e menos com dados. Com seu conhecimento teórico a respeito da dinâmica informacional, esse é um espaço que o profissional da informação pode ocupar na ciência de dados, debatendo e colaborando com a construção de modelos para análise de dados.

Além disso, outro ponto que mostrar-se relevante para os conhecimentos de um profissional da informação é a temática qualidade dos dados e informações em ambientes *big data*. Como defendido por Fagundes, Macedo e Freund (2018, p. 206), o profissional da

informação pode contribuir com a aplicação de conceitos sobre dados e informações que vão além do paradigma computacional, como por exemplo, por meio da gestão e fluxo dos dados.

Portanto, apesar da temática *big data*, ciência de dados e cientista de dados ainda ser vista quase que exclusivamente com o olhar das tecnologias, da ciência da computação e/ou da estatística, áreas, de fato, extremamente importantes para o assunto. Todavia, existe, sim, espaço para debate e colaboração dentro de outros campos, inclusive, dentro da CI. Isso se justifica pelo espaço que a disciplina de CI passou a ocupar, se preocupando com a compreensão do ciclo que a informação percorre, posto que a finalidade máxima do uso de grandes volumes de dados é justamente a geração de informação que possibilitará vantagem competitiva para as organizações, ou geração de novos conhecimentos para as diversas ciências. Desta forma, fica como sugestão para pesquisas futuras, o detalhamento, por meio de teorias e modelos, de como o profissional da informação pode contribuir com o ambiente *big data* e a ciência de dados, assim como o maior detalhamento de quais disciplinas o curso de biblioteconomia da UnB poderia agregar ao seu currículo para auxiliar na formação do profissional bibliotecário em relação aos aspectos da ciência de dados.

REFERÊNCIAS

ANGELONI, Maria Terezinha. Elementos intervenientes na tomada de decisão. *Ciência da Informação*, Brasília, v. 32, n. 1, p. 17-22, jan./abr. 2003. Disponível em: <http://revista.ibict.br/ciinf/article/view/1015/1070>. Acesso em: 3 mar. 2018.

BELLUZZO, Regina Célia Baptista. As competências do profissional da informação nas organizações contemporâneas. *RBBD. Revista Brasileira de Biblioteconomia e Documentação*, São Paulo, v. 7, n. 1, p. 58-73, jan./jun. 2011. Disponível em: <https://rbbd.febab.org.br/rbbd/article/view/180/186>. Acesso em: 26 abr. 2018.

BORKO, H. Information science: what is it?. *American Documentation*, Baltimore, v. 19, n. 1, p. 3-5, Jan. 1968. Disponível em: <https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/k---artigo-01.pdf>. Acesso em: 19 fev. 2018.

BUCKLAND, M. K. Information as thing. *Journal of the American Society for Information Science*, [S.l.], v. 45, n. 5, p. 351-360, June 1991. Disponível em: <https://pdfs.semanticscholar.org/b3d4/d7980d6a628b503003ef4e7763a93544508e.pdf>. Acesso em: 3 mar. 2018.

CAPURRO, Rafael; HJORLAND, Birger. O conceito de informação. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 12, n. 1, p. 148-207, jan./abr. 2007. Disponível em: <http://bogliolo.eci.ufmg.br/downloads/CAPURRO.pdf>. Acesso em: 3 mar. 2018.

CHEN, Hsinchun; CHIANG, Roger H. L.; STOREY, Veda C. Business intelligence and analytics: from big data to big impact. *MIS Quartely*, [S.l.], v. 36, n. 4, p. 1.165-1.188, Dec. 2012. Disponível em: <https://ai.arizona.edu/sites/ai/files/MIS611D/chen-bi-december-2012.pdf>. Acesso em: 12 abr. 2018.

CONEGLIAN, Caio Saraiva; GONÇALVEZ, Paula Regina Ventura Amorim; SANTARÉM SEGUNDO, José Eduardo. O professional da informação na era do *big data*. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, Florianópolis, v. 22, n. 50, p. 128-143, set./dez. 2017. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v22n50p128>. Acesso em: 15 mar. 2018.

CRAWFORD, Kate; MILTNER, Kate; GRAY, Mary L. Critiquing big data: politics, ethics, epistemology. *International Journal of Communication*, [S.l.], v. 8, p. 1.663-1.672, 2014. Disponível em: <http://ijoc.org/index.php/ijoc/article/view/2167/1164>. Acesso em: 12 abr. 2018.

CUNHA, Miriam Vieira da. O professional da informação e o mercado de trabalho. *Informação & Sociedade: Estudos*, João Pessoa, v. 10, n. 1, p. 1-5, 2000. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/347/269>. Acesso em: 13 mar. 2018.

CURTY, Renata Gonçalves; SERAFIM, Jucenir da Silva. A formação em ciência de dados: uma análise preliminar do panorama estadunidense. *Informação & Informação*, Londrina, v. 21, n. 2, p. 307-328, maio/ago. 2016. Disponível em:

<http://www.uel.br/revistas/uel/index.php/informacao/article/view/27945>. Acesso em: 21 mar. 2018.

DAVENPORT, Thomas H. *Big data no trabalho*: derrubando mitos e descobrindo oportunidades. Tradução de Cristina Yamagami. Rio de Janeiro: Elsevier, 2014.

DAVENPORT, Thomas H.; PATIL, D. J. Data scientists: the sexiest job of the 21st century. *Harvard Business Review*, [S.l.], v. 90, n. 10, p. 70-76, Oct. 2012.

DINO. Ciência de dados – *big data*, o primeiro curso de graduação do Brasil. *Exame*, São Paulo, 27 fev. 2018. Disponível em: <https://exame.abril.com.br/negocios/dino/ciencia-de-dados-big-data-o-primeiro-curso-de-graduacao-do-brasil/>. Acesso em: 6 abr. 2018.

FAGUNDES, Priscila Basto; MACEDO, Douglas Dyllon Jeronimo de; DUTRA, Moisés Lima. Uma análise das relações entre a qualidade da informação e *big data*. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. *Anais*... Marília: Unesp, 2017. p. [1-16]. Disponível em: <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/107/837>. Acesso em: 12 jul. 2018.

FAGUNDES, Priscila Basto; MACEDO, Douglas Dyllon Jeronimo de; FREUND, Gislaine Parra. A produção científica sobre qualidade de dados em *big data*: um estudo na base de dados *Web of Science*. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, Campinas, v. 16, n. 1, p. 194-210, jan./abr. 2018. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8650412>. Acesso em: 12 jul. 2018.

FERREIRA, Danielle Thiago. Profissional da informação: perfil de habilidades demandadas pelo mercado de trabalho. *Ciência da Informação*, Brasília, v. 32, n. 1, p. 42-49, jan./abr. 2003. Disponível em: <http://revista.ibict.br/ciinf/article/view/1018>. Acesso em: 13 mar. 2018.

FINZER, William. The data science education dilemma. *Technology Innovations in Statistics Education*, California, v. 7, n. 2, p. 1-9, 2013. Disponível em: <http://escholarship.org/uc/item/7gv0q9dc>. Acesso em: 21 mar. 2018.

FIRMANI, Donatella et al. On the Meaningfulness of "big data quality". *Data Science and Engineering*, [S.l.], v. 1, n. 1, p. 6-20, Mar. 2016. Disponível em: <https://link.springer.com/article/10.1007/s41019-015-0004-7>. Acesso em: 12 jul. 2018.

FRICKÉ, Martin. Big data and its epistemology. *Journal of the Association for Information Science and Technology*, [S.l.], v. 66, n. 4, p. 651-661, 2015.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management*, [S.l.], v. 35, n. 2, p. 137-144, Apr. 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401214001066?via%3Dihub>. Acesso em: 12 abr. 2018.

GIL, Antonio Carlos. *Como elaborar projetos de pesquisa*. 4. ed. São Paulo: Atlas, 2002.

GUIMARÃES, José Augusto Chaves. Moderno profissional da informação: elementos para sua formação no Brasil. *Transinformação*, Campinas, v. 9, n. 1, p. 124-137, jan./abr. 1997. Disponível em: <http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/1597>. Acesso em: 24 abr. 2018.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data mining*: concepts and techniques. 3. ed. Amsterdam: Elsevier, 2012.

KIM, John Yohahn; LEE, Choong Kwon. An empirical analysis of requirements for data scientists using online job posting. *International Journal of Software Engineering and Its Applications*, [S.l.], v. 10, n. 4, p. 161-172, 2016. Disponível em: <http://www.sersc.org/journals/IJSEIA/vol10_no4_2016/15.pdf>. Acesso em: 5 abr. 2018.

LANEY, Doug. 3D data management: controlling data volume, velocity, and variety. *Meta Group*, [S.l.], 6 Feb. 2001. Disponível em: <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Acesso em: 10 abr. 2018.

LE COADIC, Yves-François. *A ciência da informação*. Tradução de Maria Yêda F. S. de Figueiras Gomes. Brasília: Briquet de Lemos, 1996.

LOUKIDES, Mike. What is data science?: the future belongs to the companies and people that turn data into products. *O'Reilly Radar*, [S.l.], June 2, 2010.

MANETH, Sebastian; POULOVASSILIS, Alexandra. Data science. *The Computer Journal*, [S.l.], v. 60, n. 3, p. 285-286, 2017.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. *Fundamentos de metodologia científica*. 5. ed. São Paulo: Atlas, 2003.

MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. *Big data*: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Tradução de Paulo Polzonoff Junior. Rio de Janeiro: Elsevier, 2013.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big data: the management revolution. *Harvard Business Review*, [S.l.], v. 90, n. 10, p. 60-68, Oct. 2012. Disponível em: <https://ai.arizona.edu/sites/ai/files/MIS510/big_data_-_the_management_revolution_0.pdf>. Acesso em: 12 abr. 2018.

MILLER, Claire Cain. Data science: the numbers of our lives. *The New York Times*, New York, 23 Apr. 2013. Disponível em: <https://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html>. Acesso em: 28 mar. 2018.

ORTEGA, Cristina Dota. Relações históricas entre biblioteconomia, documentação e ciência da informação. *DataGramaZero – Revista de Ciência da Informação*, Rio de Janeiro, v. 5, n. 5, out. 2004. Não paginado. Disponível em: <http://www.brapci.inf.br/index.php/article/view/0000002048/e908b9a74b0fb8f5aff3bd1881eec6b2/>. Acesso em: 10 mar. 2018.

PINHEIRO, Lena Vania Ribeiro; LOUREIRO, José Mauro Matheus. Traçados e limites da ciência da informação. *Ciência da Informação*, Brasília, v. 24, n. 1, abr. 1995. Não paginado. Disponível em: <http://revista.ibict.br/ciinf/article/view/609>. Acesso em: 6 mar. 2018.

PORTO, Fábio; ZIVIANI, Artur. Ciência de dados. In: SEMINÁRIO DE GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL, 3., 2014, Rio de Janeiro. *Anais*... Rio de Janeiro: [SBC], 2014. p. [1-18]. Disponível em: <http://www.lncc.br/~ziviani/papers/III-Desafios-SBC2014-CiD.pdf>. Acesso em: 19 dez. 2017.

PRESS, Gil. A very short story of data science. *Forbes*, [S.l.], May 28, 2013. Disponível em: <http://www.mat.ufrgs.br/~viali/estatistica/mat2274/material/textos/A%20Very%20Short%20History%20Of%20Data%20Science%20-%20Forbes.pdf>. Acesso em: 19 dez. 2017.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. *Metodologia do trabalho científico*: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013.

RIBEIRO, Claudio José Silva. *Big data*: os novos desafios para o profissional da informação. *Informação & Tecnologia (ITEC)*, João Pessoa/Marília, v. 1, n. 1, p. 96-105, jan./jun. 2014. Disponível em: <http://periodicos.ufpb.br/ojs/index.php/itec/article/view/19380/11156>. Acesso em: 4 jan. 2018.

RODRIGUES, Adriana Alves; NÓBREGA, Emeide; DIAS, Guilherme Ataíde. Desafios da gestão de dados na era *big data*: perspectivas profissionais. In: ENCONTRO NACIONAL EM PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. *Anais*... Marília, SP: Unesp, 2017. p. [1-19]. Disponível em: <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/376/862>. Acesso em: 5 abr. 2018.

SANTA ANNA, Jorge. O moderno profissional da informação à luz dos paradigmas da ciência da informação. *Biblionline*, João Pessoa, v. 13, n. 2, p. 18-29, abr./jun. 2017. Disponível em: <http://www.periodicos.ufpb.br/ojs2/index.php/biblio/article/view/32501/18818>. Acesso em: 24 abr. 2018.

SANTA ANNA, Jorge; PEREIRA, Gleice; CAMPOS, Suelen de Oliveira. Sociedade da informação x biblioteconomia: em busca do moderno profissional da informação (MIP). *RBBD. Revista Brasileira de Biblioteconomia e Documentação*, São Paulo, v. 10, n. 1, p. 68-85, jan./jun. 2014. Disponível em: <https://rbbd.febab.org.br/rbbd/article/view/293/293>. Acesso em: 13 mar. 2018.

SANTOS, Jussara Pereira. O moderno profissional da informação: o bibliotecário e seu perfil face aos novos tempos. *Informação & Informação*, Londrina, v. 1, n. 1, p. 5-13, jan./jun. 1996. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/viewFile/1613/1367>. Acesso em: 24 abr. 2018.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em:

<http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235/22>. Acesso em: 5 jan. 2018.

SIQUEIRA, Jéssica Câmara. Biblioteconomia, documentação e ciência da informação: história, sociedade, tecnologia e pós-modernidade. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 15, n. 3, p. 52-66, set./dez. 2010. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/1124/771>. Acesso em: 10 mar. 2018.

TARGINO, Maria das Graças. Quem é o professional da informação?. *Transinformação*, Campinas, v. 12, n. 2, p. 61-69, jul./dez. 2000. Disponível em: <http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/1531/1505>. Acesso em: 14 mar. 2018.

VALENTIM, Marta Lígia Pomim. O moderno profissional da informação: formação e perspectiva profissional. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, Florianópolis, v. 5, n. 9, p. 16-28, jan. 2000. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2000v5n9p16>. Acesso em: 24 abr. 2018.

VICTORINO, Márcio de Carvalho et al. Uma proposta de ecossistema de *big data* para a análise de dados abertos governamentais conectados. *Informação & Sociedade: Estudos*, João Pessoa, v. 27, n. 1, p. 225-242, jan./abr. 2017. Disponível em: <http://www.periodicos.ufpb.br/ojs2/index.php/ies/article/view/29299/17505>. Acesso em: 9 abr. 2018.

WERSIG, Gernot; NEVELING, Ulrich. The phenomena of interest to information science. *Information Scientist*, [S.l.], v. 9, n. 4, p. 127-140, Dec. 1975. Disponível em: <http://sigir.org/files/museum/pub-13/18.pdf>. Acesso em: 17 fev. 2018.

ZIVIANI, Artur; PORTO, Fábio; OGASAWARA, Eduardo. Ciência de dados: desafio para a ciência, indústria e governo. *ComCiência: Revista Eletrônica de Jornalismo Científico*, Campinas, v. 170, jul. 2015. Não paginado. Disponível em: <http://comciencia.br/comciencia/handler.php?section=8&edicao=115&id=1381>. Acesso em: 20 dez. 2017.